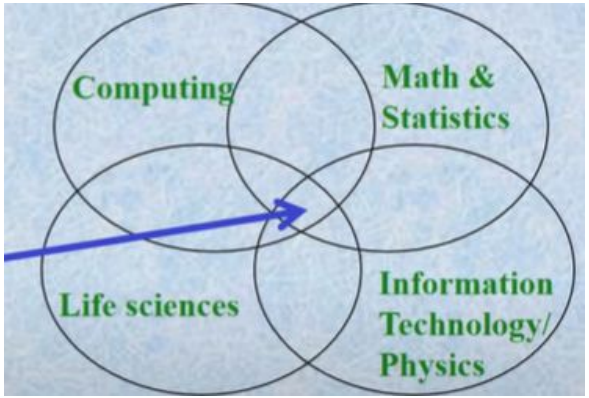
Bioinformatics

Bulbul Ahammad
Assistant Prof., Dept. of CSE, JU
Email: bulbul@juniv.edu

What is Bioinformatics?

- Bioinformatics is an interdisciplinary field of science in which biology, statistics/mathematics, computer science and information technology merge together to form a single discipline in order to analyze biological information using computers and statistical techniques.
- The term bioinformatics was coined by Paulien Hogeweg and Ben Hesper in 1970 for the study of informatics in biological systems. Margaret Oakley Dayhoff, an american physical chemist is called as the father of bioinformatics.
- Biological experiments can generate large amounts of data (macromolecular sequences, structures, expression profiles, pathways and so on.)



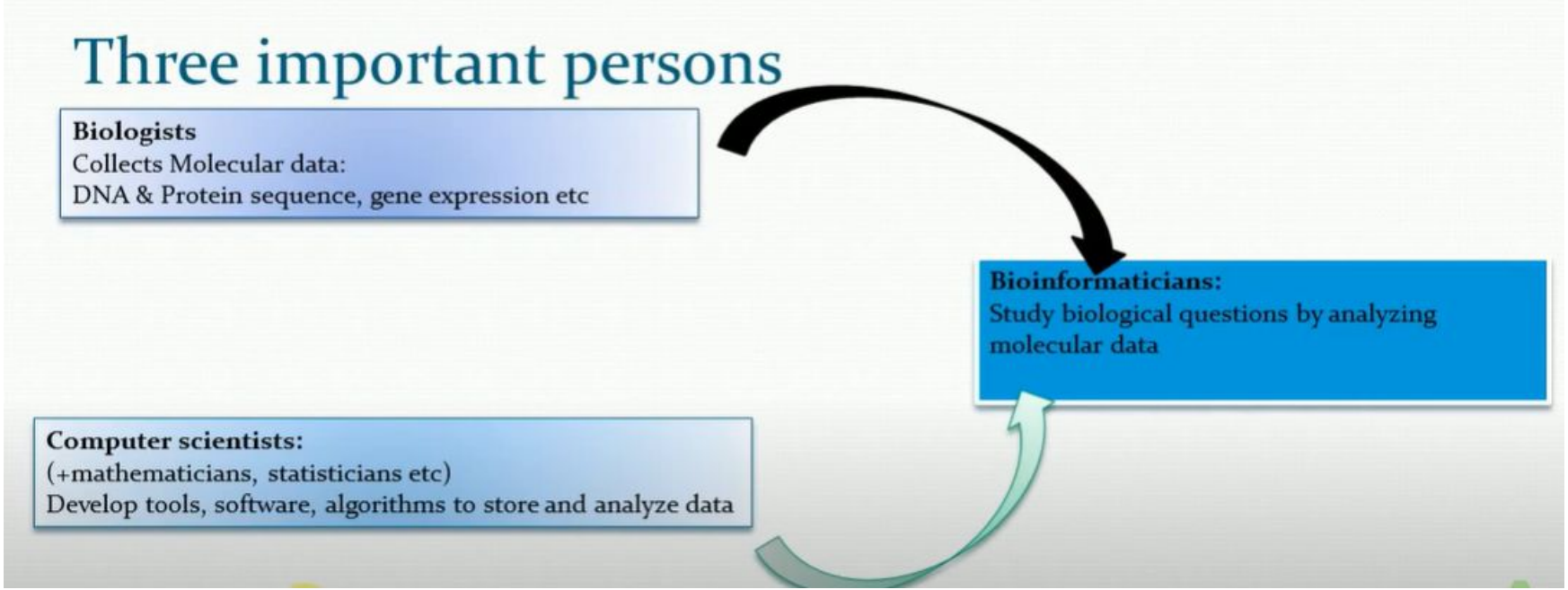
Bioinformatics is about acquiring/ collecting, managing, analyzing and understanding those data.

Also termed as biocomputing or computational Biology.

What is Bioinformatics?

Everyone of us have a book of fate in us. In which everything about us is written. For example: what would be our gender, the height of which we would grow, the skin color, which diseases we are immune to and which diseases we are susceptible to. This is book of fate is our genome or DNA. The hugeness of our DNA has made it almost impossible to understand it completely but this is until the last decade.

What is Bioinformatics?



Bioinformatics Glossary

- **Annotation:** An annotation is extra information associated with a particular point in a document or other piece of information. It can be a note that includes a comment or explanation. Annotations are sometimes presented in the margin of book pages.

- **Organism:**
  A. Any living biological entity, such as an animal, plant, fungus, or bacterium. [Collins English Dictionary]
  B. An organism refers to a living thing that has an organized structure, can react to stimuli, reproduce, grow, adapt, and maintain homeostasis. [Biology Online Dictionary]

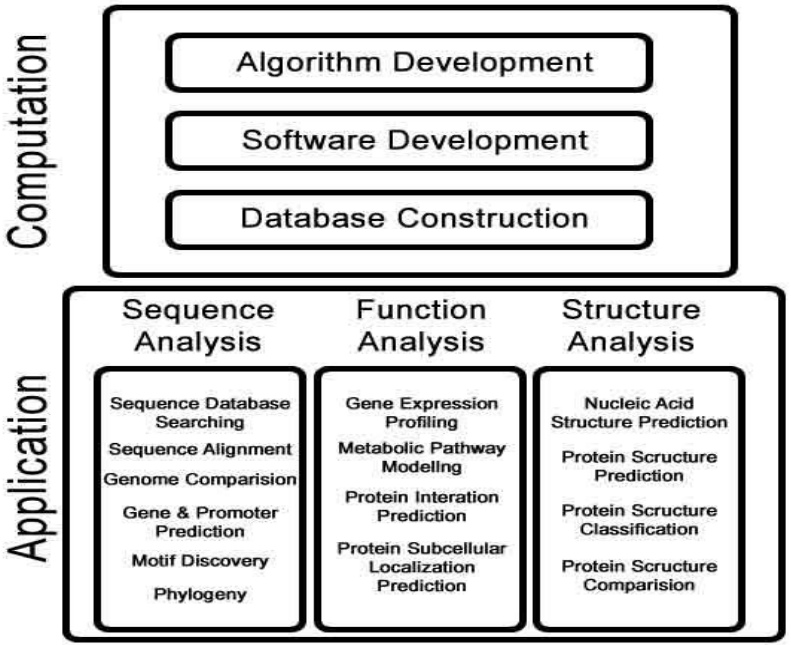
-

Goals of Bioinformatics

To construct and manage biological database.

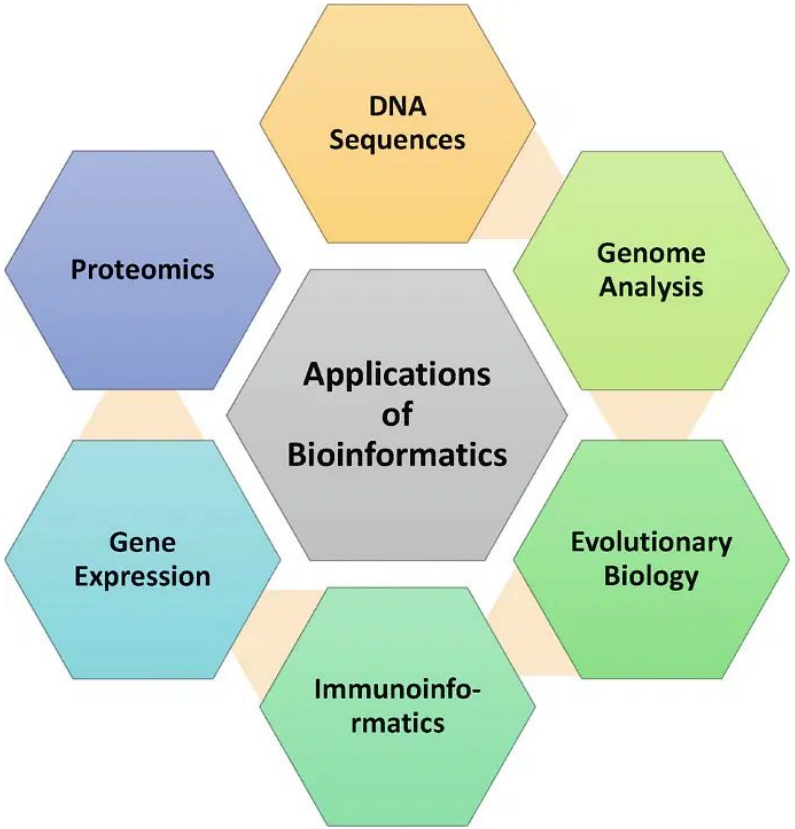
To develop algorithm for different types of sequence analysis.

To develop softwares.



Applications of Bioinformatics

Bioinformatics and its application depend on taking out useful facts and figures from a collection of data reserved to be processed into useful information. Bioinformatics focuses its scope on the areas of 3D image processing, 3D modeling of living cells, image analysis, drug development, and a lot more. The most important application of bioinformatics can be seen in the field of medicine, where its data is used to create antidotes for infectious and harmful diseases.



Uses of Bioinformatics

Bioinformatics is used in various fields, especially in biomedicine where it has various applications.

Drug discovery: By the use of structure-based drug design with Bioinformatics, scientists can come up with effective drugs for various illnesses and acute diseases.

Personalized medicine: By assessing the genetic structure of the patient and their respective medical history, personalized medicine can be curated which would prove to be much more effective.

Preventive medicines: This is majorly done by integrating data from Bioinformatics, Bioanalytics, and Epidemiology. As the name says, preventive medicines can prevent a chain of transmission of diseases or mitigate any acute disease before its onset.

Uses of Bioinformatics

Bioinformatics is used in various fields, especially in biomedicine where it has various applications.

Gene therapy: Gene therapy is the process where defective genes are replaced by new ones in the gene structure of a living organism. Since each organisms gene structure is very different, huge data may be involved to produce an accurate replacement.

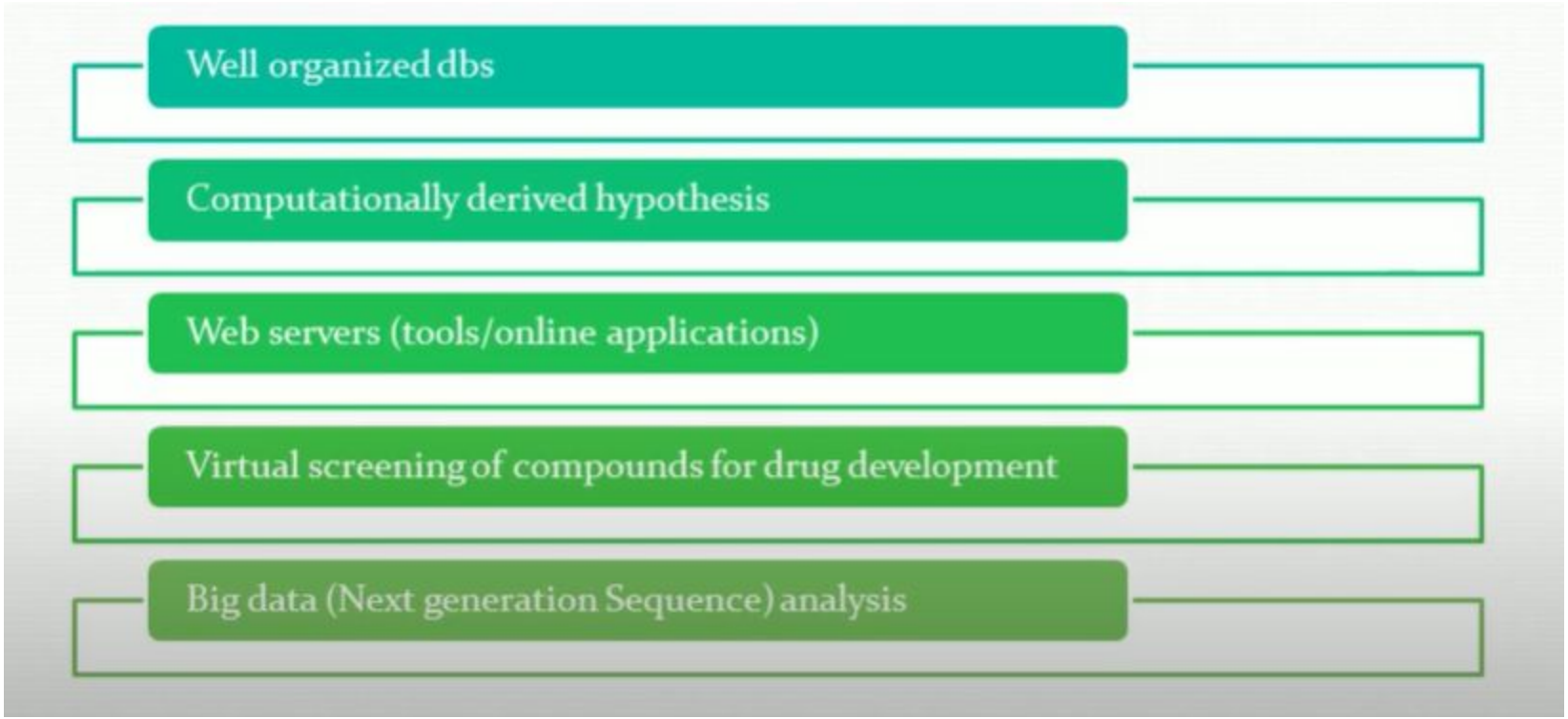
Fields Dealt by Bioinformatics (Don't read this slide)

Fields that bioinformatics deals with:

1. **Genomics**
2. **Sequencing**
3. **Personalized drug design**
4. **Drug discovery**
5. **AL and ML in biology**

Personalized drug design: From the beginning of the civilization, drug has been used to cure diseases. Conventionally, a single drug is administered to everybody even if we know that the body of every individual respond differently to the same drug. It may surely cure the targeted disease but may also induce site reaction in the body of the individual. Now, somewhat personalized drugs are used that is also according to the compatibility of the individual body to the drug but it is not the true personalization of the medicine. Bioinformatics has taken the concept of personalized medicine to a whole new level. The personalised medicine can now be designed at the genetic level. That means, a drug can be designed for a particular individual and has no side effect for that individual.

Major Aspects of Bioinformatics



Areas Where Bioinformatics is Applied!

- **Genomics:**
  - ★ Sequencing data analysis
  - ★ Genomic feature prediction
- **Proteomics:**
  - ★ Protein 3D structure modeling
  - ★ Drug design
- **System Biology:**
  - ★ Genome set enrichment
  - ★ Pathway analysis
- **Phenotype:**
  - ★ Image analysis
  - ★ Integration

What is the purpose of Bioinformatics?

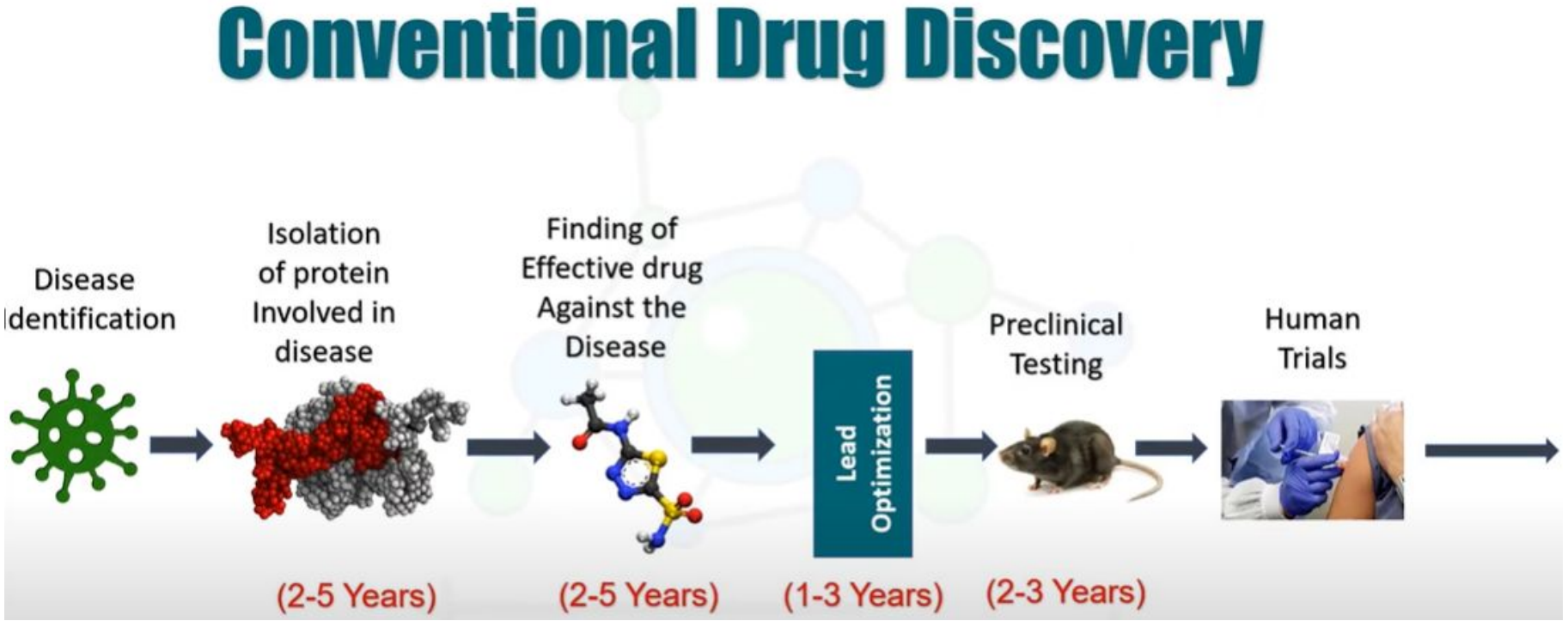
- **The work of bioinformatics is in the processing of large amount of data.**
- **The purpose of bioinformatics is to make sense of processes taking place inside organism or living systems.**

How is Bioinformatics Different from Conventional Biology

- **Drawbacks of conventional biology**
  - ★ **High resource requirement: conventional research generally requires a sophisticated lab to work in and may also need expansive utilities and model organisms.**
  - ★ **The experiments are normally long, delicate and sometimes risky to the researchers**

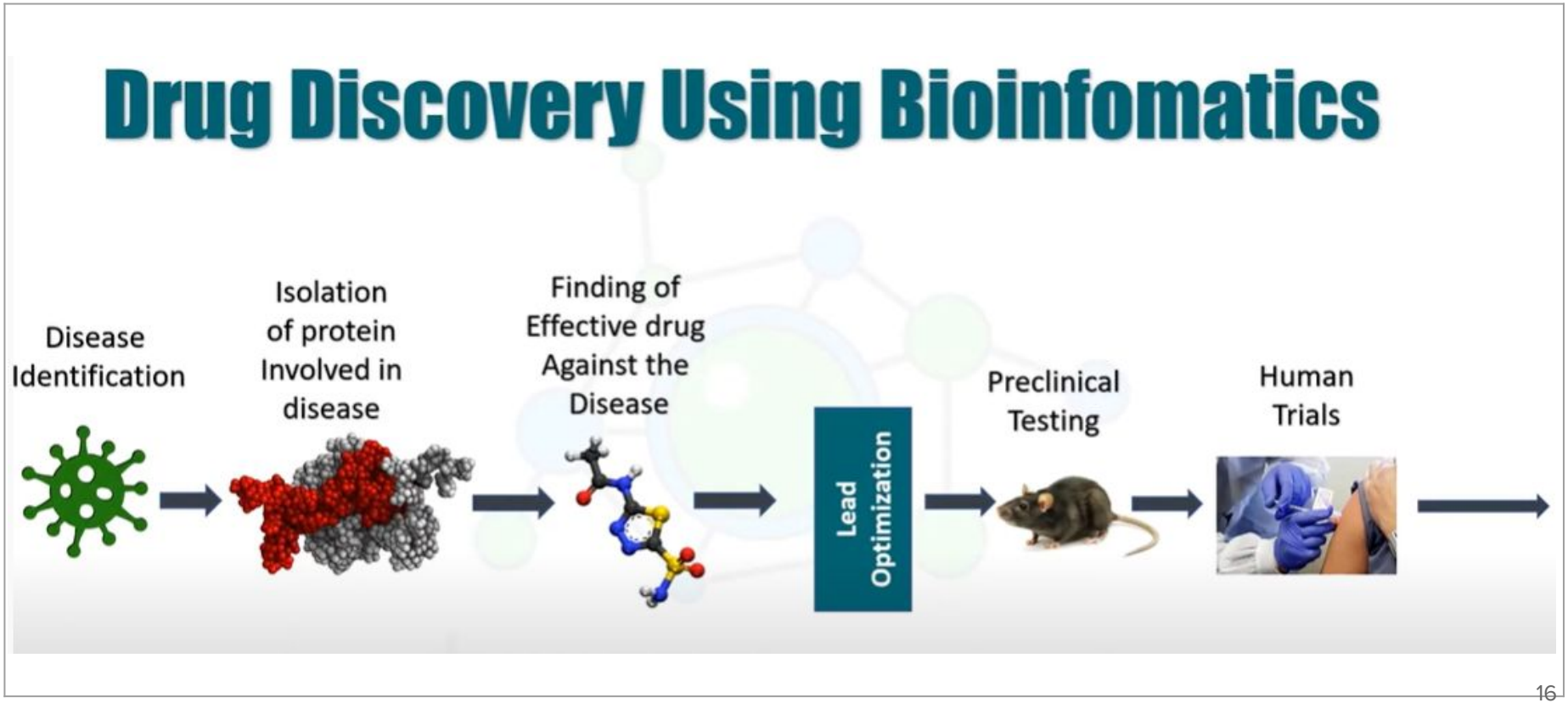
Let's take an example of drug discovery of a disease and see it in both ways: conventional biological approach and bioinformatic approach.

How is Bioinformatics Different from Conventional Biology



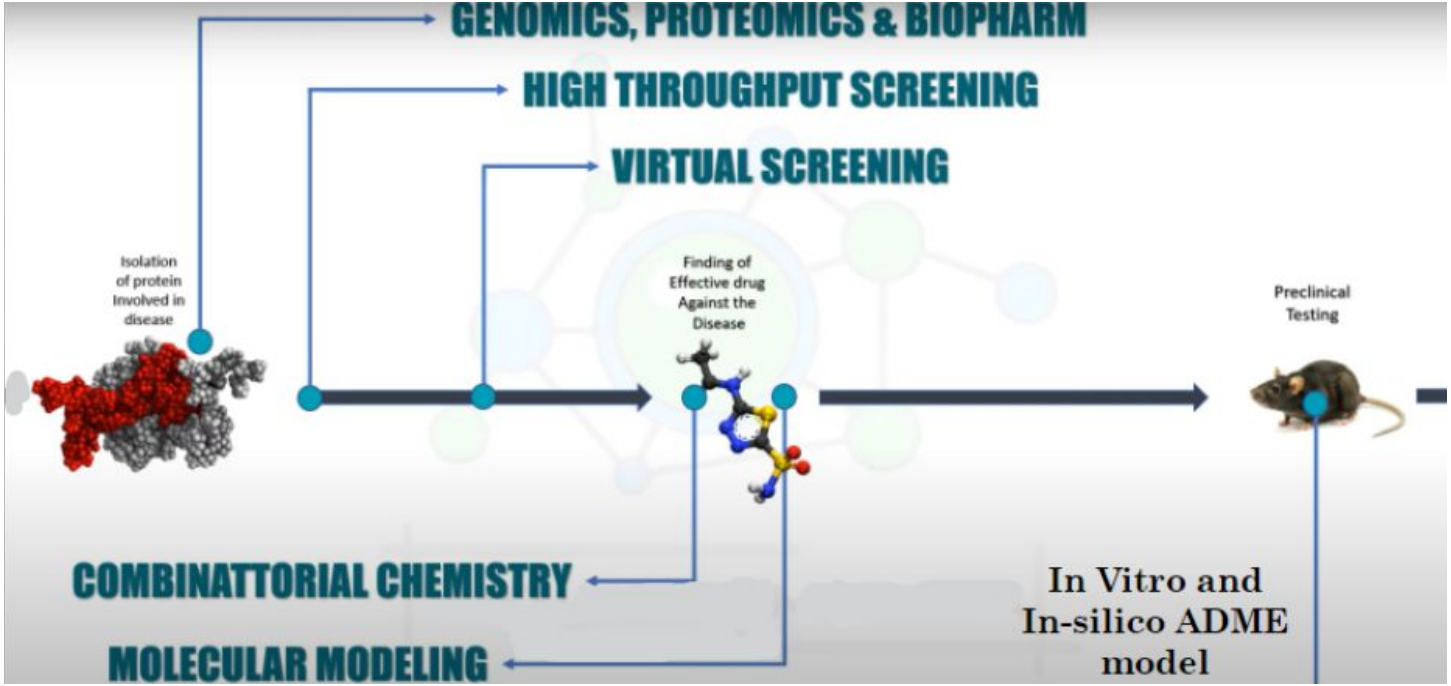
Conventional approach is lengthy, expansive and risky with high failure rate.

How is Bioinformatics Different from Conventional Biology



How is Bioinformatics Different from Conventional Biology

Almost all the steps are done in in-silico environment or using computers which itself cut the expenses of materials and other stuffs. Pre-clinical testing is done on model organism without harming any real test organism.



And modern computers are capable of performing experiments within seconds allowing generation after generation experiments within a small span of time. Bioinformatics allows to produce an effective drug within a short period of time with less expense and a greater accuracy.

Biological Database

These are the databases consisting of biological data like protein sequencing, molecular structure, DNA sequences, etc in an organized form.

Several computer tools are there to manipulate the biological data like an update, delete, insert, etc. Scientists, researchers from all over the world enter their experiment data and results in a biological database so that it is available to a wider audience.

Biological databases are free to use and contain a huge collection of a variety of biological data.

Uses of Biological Database

It helps the researchers to study the available data and form a new thesis, anti-virus, helpful bacteria, medicines, etc.

It helps scientists to understand the concepts of biological phenomena.

Secondary databases have become the molecular biologist's reference library over the past decade or so, providing a wealth of information on just about any gene or gene product that has been investigated by the research community.

The database acts as a storage of biological information.

It helps remove the redundancy of data.

Types of Biological Database

There are basically three types of biological databases. They are: Primary databases, secondary databases and composite databases.

Primary Databases

It can also be called an archival database since it archives the experimental results submitted by the scientists. The primary database is populated with experimentally derived data like genome sequence, macromolecular structure, etc. The data entered here remains uncurated(no modifications are performed over the data).

It obtains unique data obtained from the laboratory and these data are made accessible to normal users without any change

Types of Biological Database

Primary Databases

The data are given accession numbers when they are entered into the database. The same data can later be retrieved using the accession number. Accession number identifies each data uniquely and it never changes.

Examples:

Examples of Primary database- Nucleic Acid Databases are GenBank and DDBJ
Protein Databases are PDB,SwissProt,PIR,TrEMBL,Metacyc, etc.

Types of Biological Database

Secondary Databases

The data stored in these types of databases are the analyzed result of the primary database. Computational algorithms are applied to the primary database and meaningful and informative data is stored inside the secondary database.

The data here are highly curated(processing the data before it is presented in the database). A secondary database is better and contains more valuable knowledge compared to the primary database.

Examples:
InterPro (protein families, motifs, and domains)
UniProt Knowledgebase (sequence and functional information on proteins)

Types of Biological Database

Composite Databases

The data entered in these types of databases are first compared and then filtered based on desired criteria.

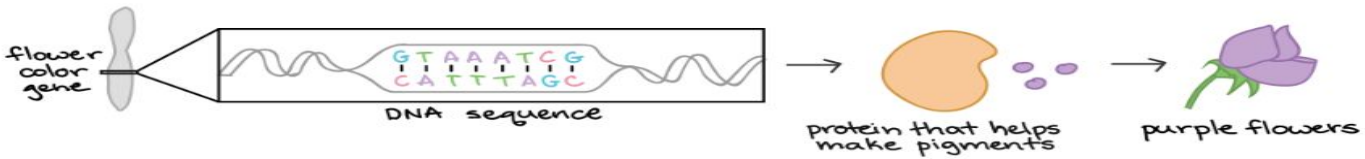
The initial data are taken from the primary database, and then they are merged together based on certain conditions. It helps in searching sequences rapidly. Composite Databases contain non-redundant data.

Examples:
Composite Databases -OWL,NRD and Swissport +TREMBL

Central Dogma of Molecular Biology

The bone, skin, and muscle that we see are made up of cells. And each of these cells contains many millions of proteins. As a matter of fact, proteins are the key molecular building blocks for every organism on earth. **But how are these proteins made in a cell!!!**

The instructions for making proteins are written in a cell's DNA. A DNA is not just a long string of nucleotides but is divided up into functional units called genes. Each gene provides instructions for a functional product, that is, a molecule needed to perform a job in a cell. In many cases, the functional product of a gene is a protein. For example, Mendel's flower color gene provides instructions for a protein that helps make colored molecules (pigments) in flower petals.



Central Dogma of Molecular Biology

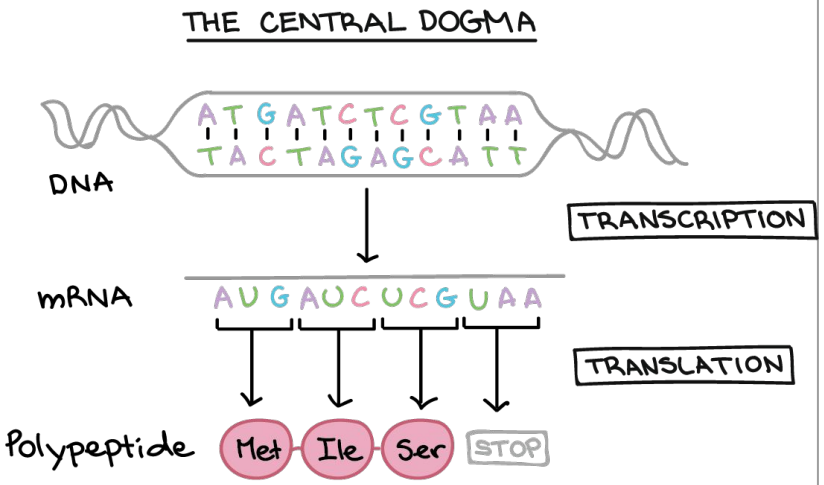
The central dogma states that the pattern of information that occurs most frequently in our cells is:

- From existing DNA to make new DNA (DNA replication)
- From DNA to make messenger RNA (Transcription)
- From mRNA to make new proteins (Translation)

In **transcription**, the DNA sequence of a gene is copied to make an RNA molecule. This step is called transcription because it involves rewriting, or transcribing, the DNA sequence in a similar RNA "alphabet." In eukaryotes, the RNA molecule must undergo processing to become a mature messenger RNA (mRNA).

Central Dogma of Molecular Biology

In **translation**, the sequence of the mRNA is decoded to specify the amino acid sequence of a polypeptide. The name translation reflects that the nucleotide sequence of the mRNA sequence must be translated into the completely different "language" of amino acids.



Thus, during expression of a protein-coding gene, information flows from DNA →RNA→protein. This directional flow of information is known as the central dogma of molecular biology. Non-protein-coding genes (genes that specify functional RNAs) are still transcribed to produce an RNA, but this RNA is not translated into a polypeptide. For either type of gene, the process of going from DNA to a functional product is known as gene expression.

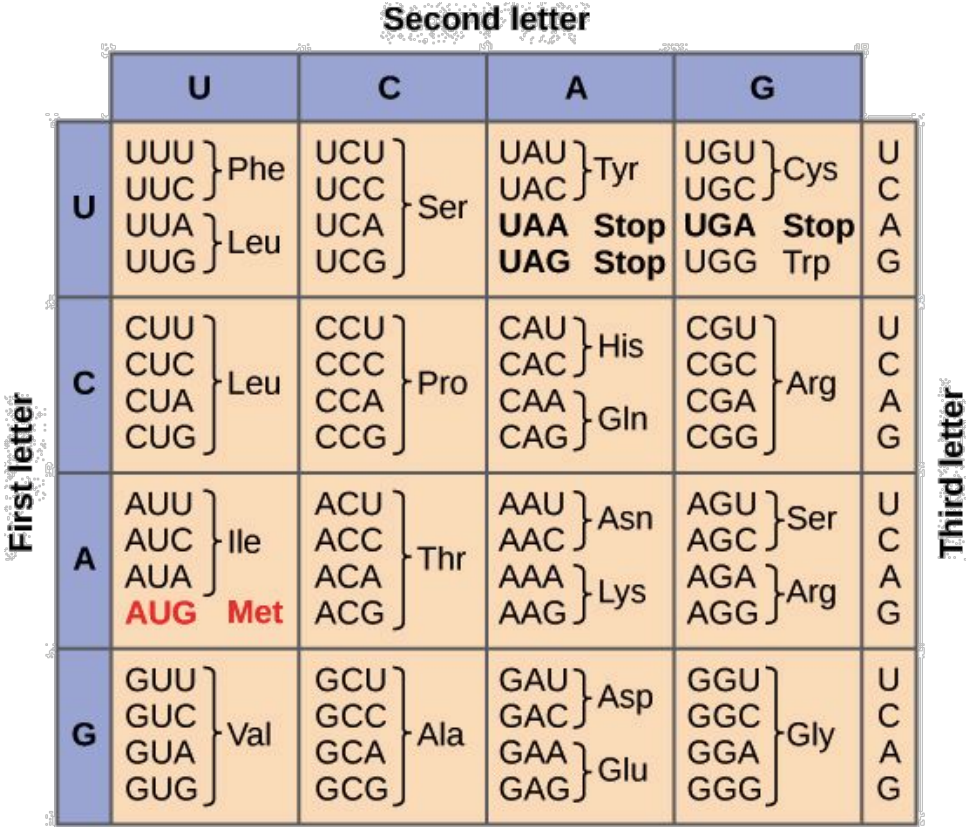
The Genetic Code

During translation, a cell “reads” the information in a messenger RNA (mRNA) and uses it to build a protein. Actually, to be a little more techical, an mRNA doesn’t always encode—provide instructions for—a whole protein. Instead, what we can confidently say is that it always encodes a polypeptide, or chain of amino acids.

In an mRNA, the instructions for building a polypeptide are RNA nucleotides (As, Us, Cs, and Gs) read in groups of three. These groups of three are called codons.

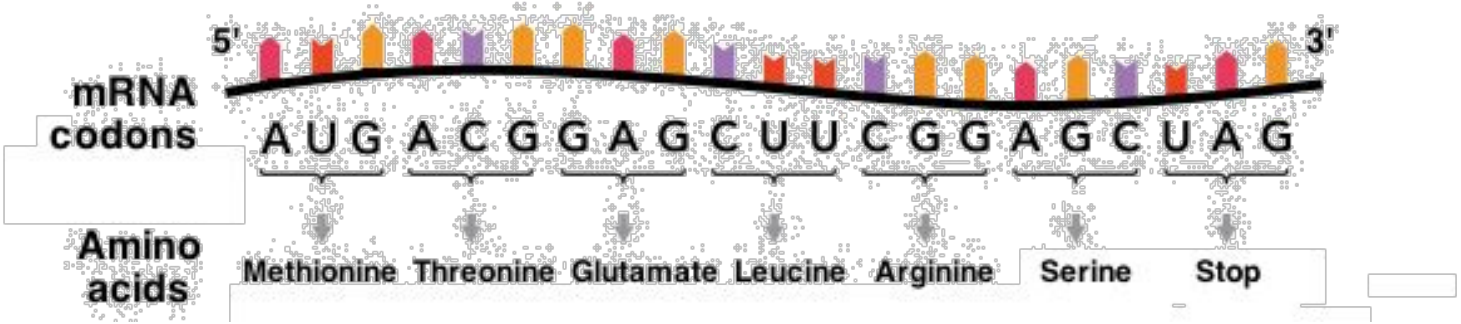
There are 61 codons for amino acids, and each of them is "read" to specify a certain amino acid out of the 20 commonly found in proteins. One codon, AUG, specifies the amino acid methionine and also acts as a start codon to signal the start of protein construction.

The Genetic Code



The Genetic Code

There are three more codons that do not specify amino acids. These stop codons, UAA, UAG, and UGA, tell the cell when a polypeptide is complete. All together, this collection of codon-amino acid relationships is called the genetic code, because it lets cells "decode" an mRNA into a chain of amino acids.



Where Do We Come From!

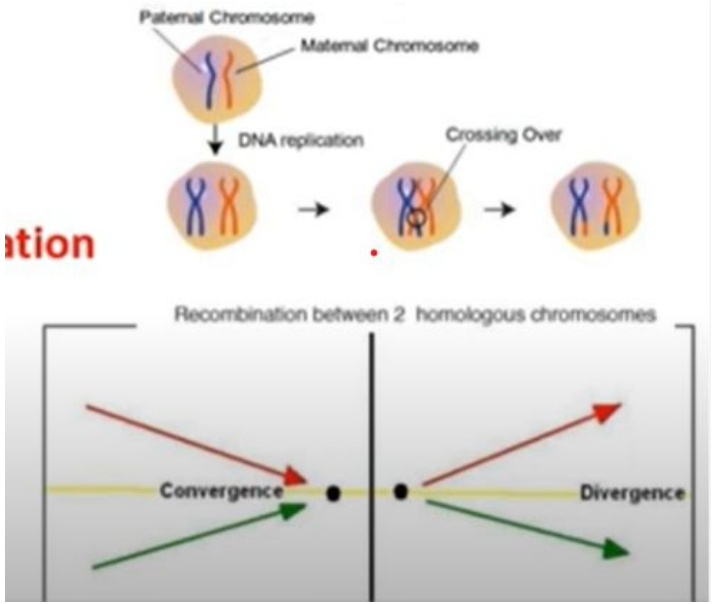
- **Regarding our coming to this world, there are different theories and hypothesis:**
  1. **Theory of creation (Creationism)**
  2. **Panspermia theory**
  3. **Theory of spontaneous generation (Abiogenesis)**
  4. **Theory of evolution (Evolutionism)**
- **People have to believe either in creationism or in evolutionism.**

Evidence of Evolution

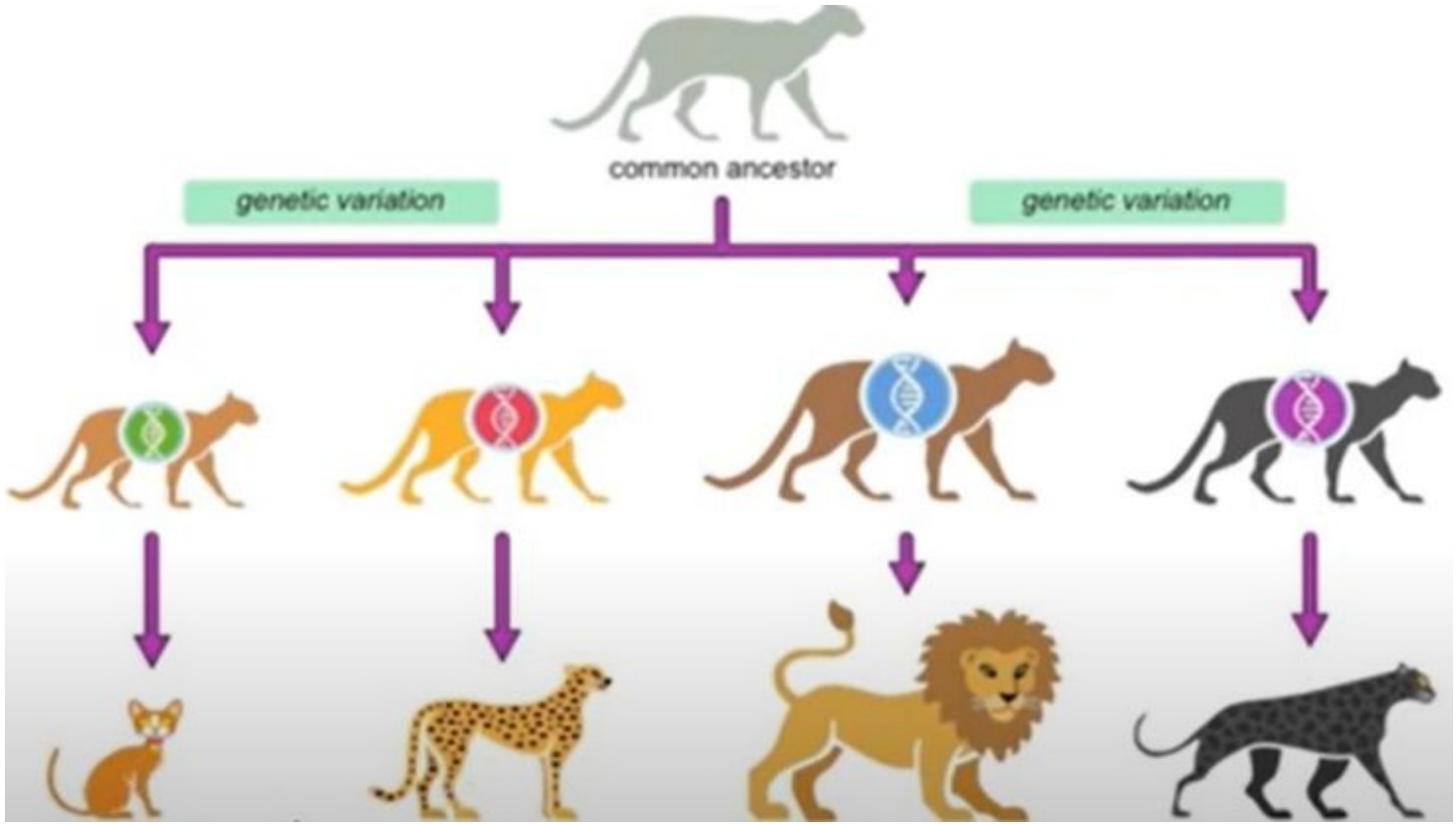
- **Development of a biological form from other preexisting forms or its origin to the current existing form.**
- **Four sources of evidences:**
  1. **Fossil Record**
  2. **Geographical Distribution of Related Species**
  3. **Homologous body structures**
  4. **Similarities in embryology**

Evidence of Evolution

- **The mechanism through which evolution occur include:**
  - **A. Natural Selection**
  - **B. Mutation**
  - **C. Sexual Recombination**
  - **D. Divergence**
  - **E. Convergence**



Evidence of Evolution



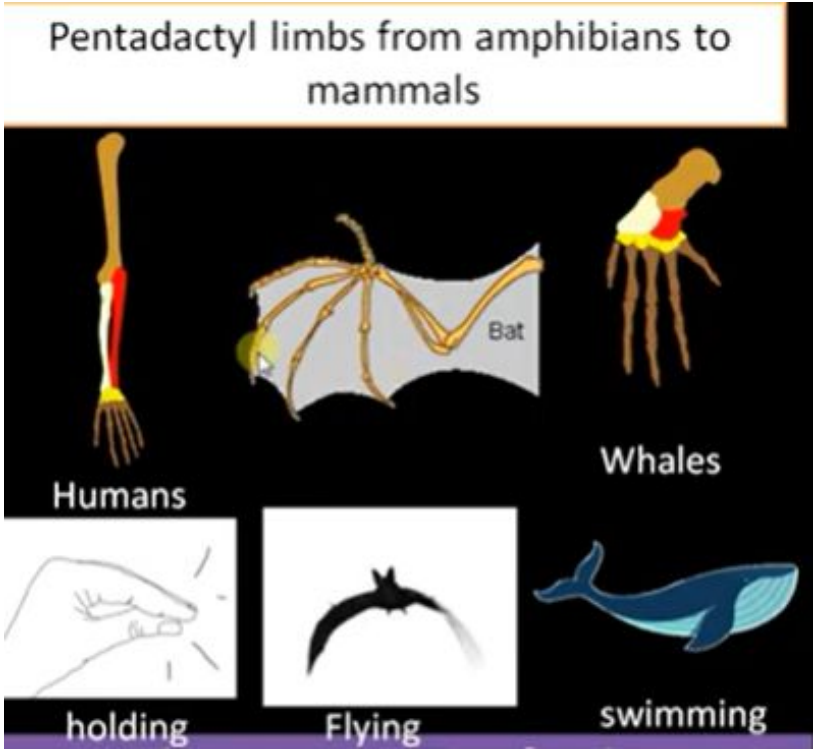
Types of Evolution

- **Evolution can be of several types:**
  1. **Divergent evolution**
  2. **Convergent evolution**

- **Speciation: Divergent evolution may lead to completely new species sometimes. This is known as speciation.**

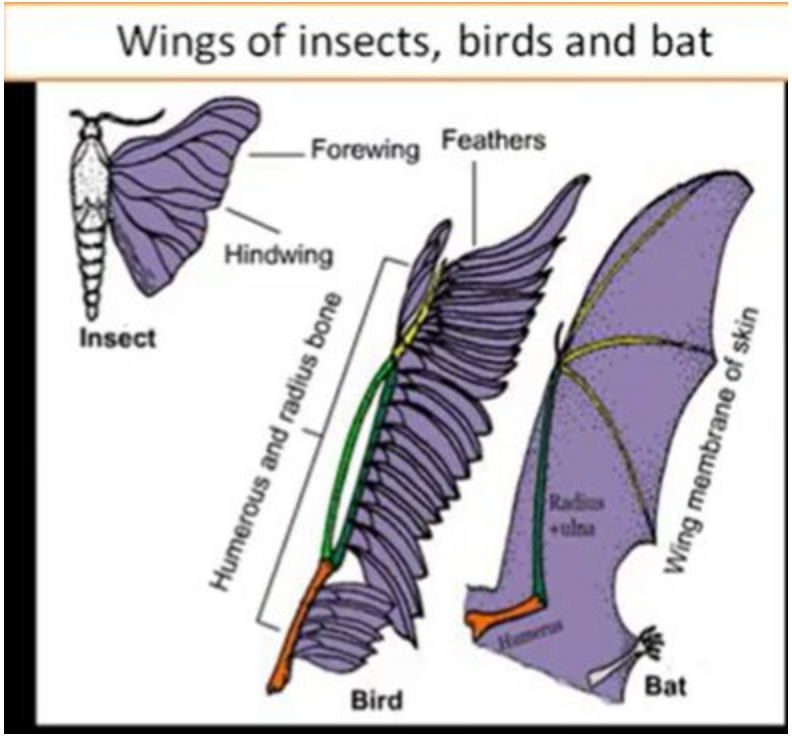
Types of Evolution (Divergent Evolution)

- **Homologous organs:** organs with similar structure (skeletal structure or anatomy) and origin, but with different function in different organisms.
- Homologous organs are developed in related animals.
- Homologous organs are inherited from a common ancestor.
- Homologous organs are due to divergent evolution.
- **Divergent evolution:** It is the process by which groups from the same common ancestor diverge or evolve and accumulate differences, resulting in the formation of new species.
- Here, human, bat and whales are mammals.



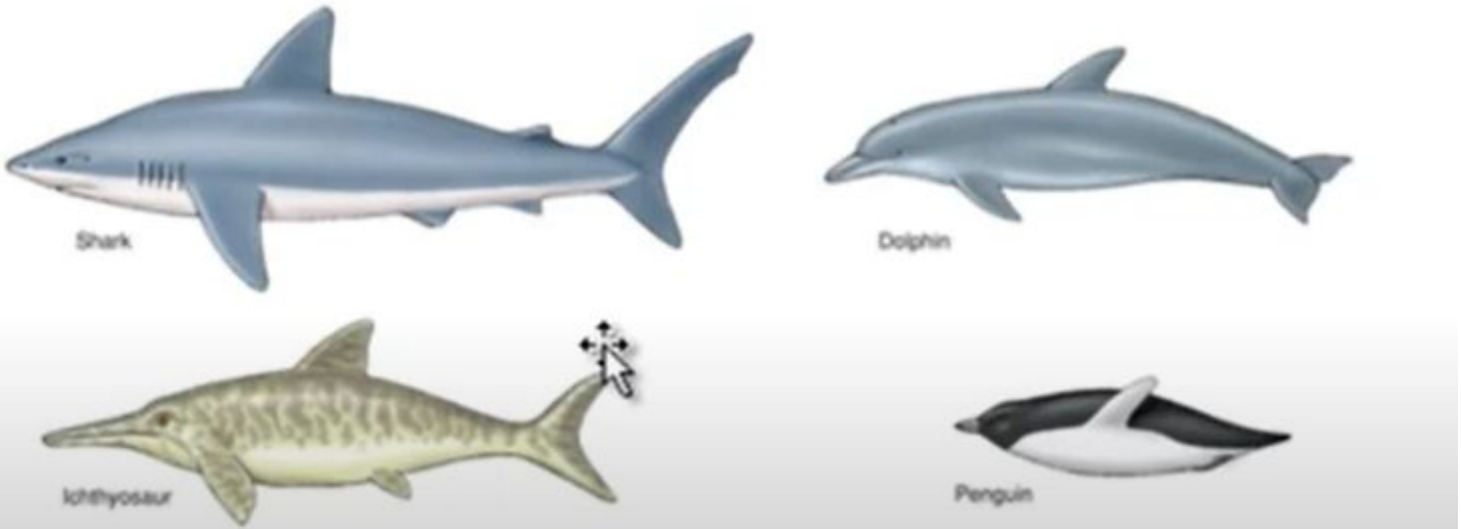
Types of Evolution (Convergent Evolution)

- **Analogous organs:** organs with different structure (skeletal structure or anatomy) and origin, but with similar function in different organisms.
- Analogous organs are developed in unrelated animals.
- Analogous organs are developed due to same selection pressure, environment or habit.
- Analogous organs are due to convergent evolution.
- **Convergent evolution:** It is the process by which unrelated groups independently evolve similar traits to adapt to similar environment.

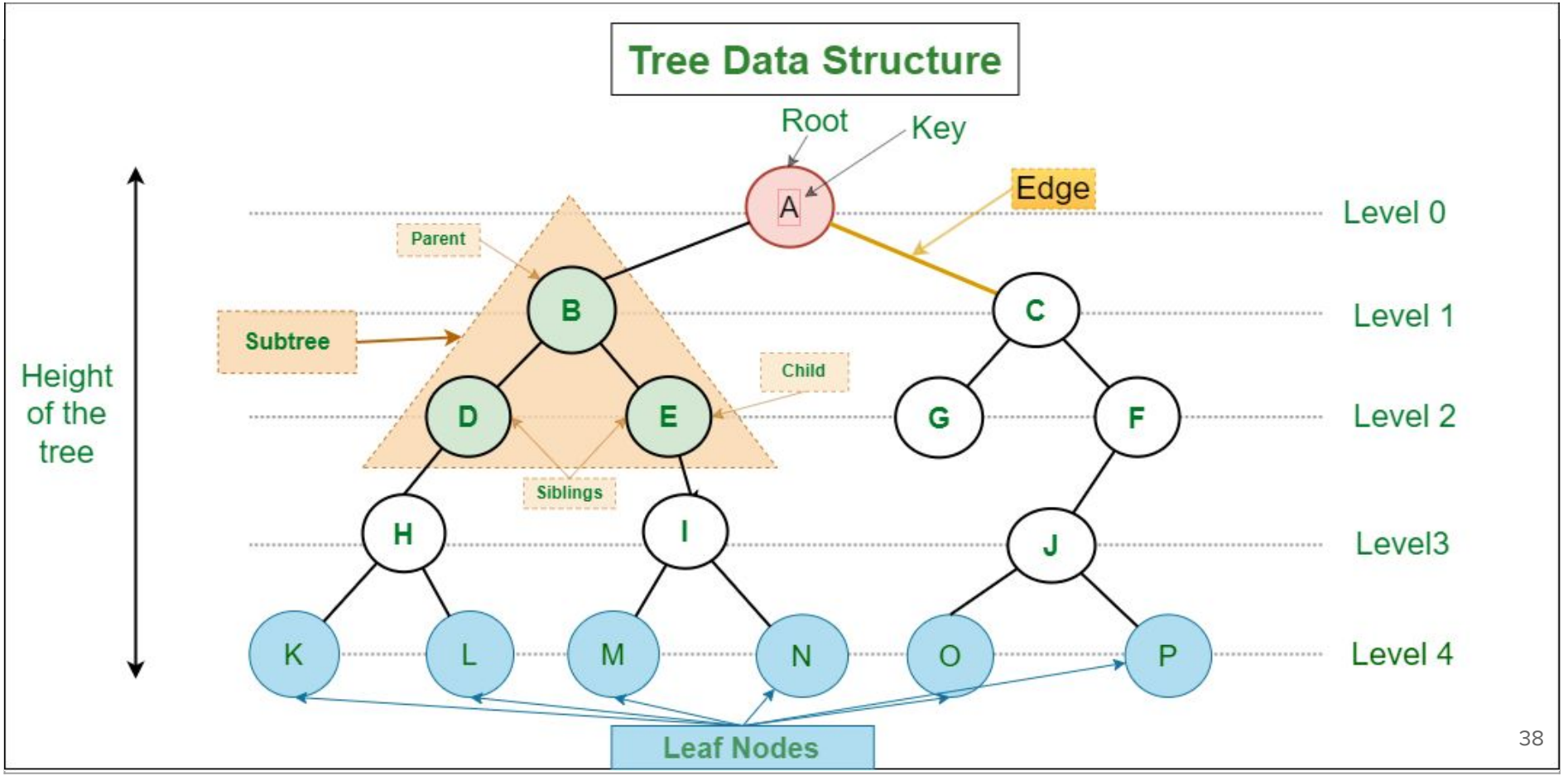


Types of Evolution (Convergent Evolution)

Convergent Evolution: Streamlining



Convergent evolution is the process by which unrelated species evolve similar physical characteristics because they have similar lifestyles


Tree Data Structure
Root
Key
Edge
A
Parent
B
C
Subtree
Child
D
E
G
F
Siblings
H
I
J
K
L
M
N
O
P
Level 0
Level 1
Level 2
Level3
Level 4
Height of the tree
Leaf Nodes

Tree Terminology

- **Parent Node:** The node which is a predecessor of a node is called the parent node of that node. **{B}** is the parent node of **{D, E}**.
- **Child Node:** The node which is the immediate successor of a node is called the child node of that node. Examples: **{D, E}** are the child nodes of **{B}.**
- **Root Node:** The topmost node of a tree or the node which does not have any parent node is called the root node. {A} is the root node of the tree. A non-empty tree must contain exactly one root node and exactly one path from the root to all other nodes of the tree.
- **Leaf Node or External Node:** The nodes which do not have any child nodes are called leaf nodes. **{K, L, M, N, O, P}** are the leaf nodes of the tree.

Tree Terminology

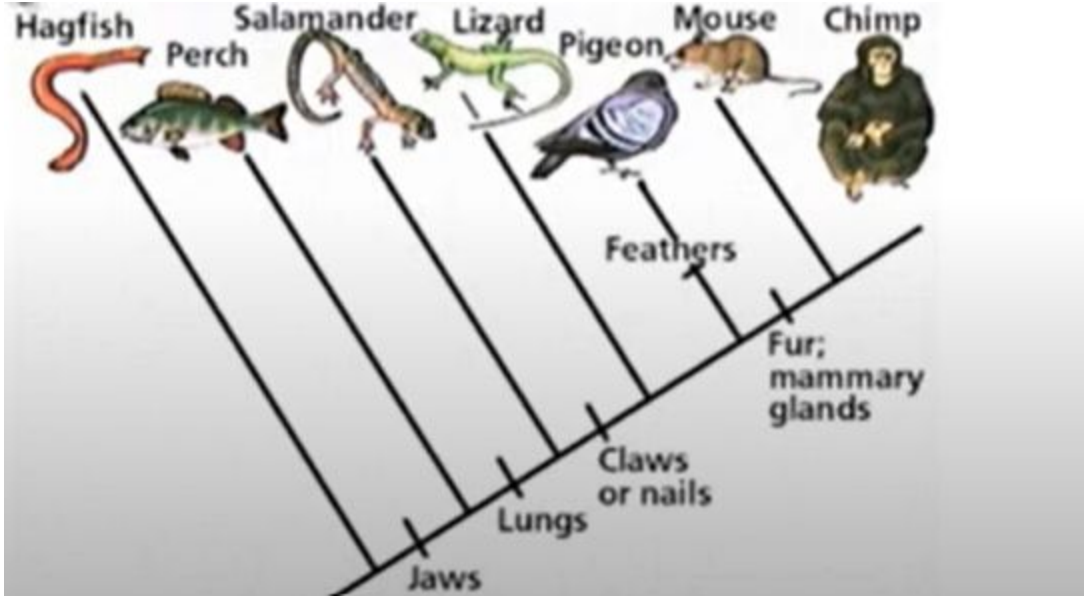
- **Ancestor of a Node:** Any predecessor nodes on the path of the root to that node are called Ancestors of that node. **{A,B}** are the ancestor nodes of the node **{E}**
- **Descendant:** Any successor node on the path from the leaf node to that node. **{E,I}** are the descendants of the node **{B}.**
- **Sibling:** Children of the same parent node are called siblings. **{D,E}** are called siblings.
- **Level of a node:** The count of edges on the path from the root node to that node. The root node has level **0**.
- **Internal node:** A node with at least one child is called Internal Node.

Phylogeny

- **Phylogeny** is the study of relationships among different groups of organisms and their evolutionary development.
- Phylogeny attempts to trace the evolutionary history of all life on the planet. It is based on the phylogenetic hypothesis that all living organisms share a common ancestry.
- The relationships among organisms are depicted in what is known as a phylogenetic tree.
- Relationships are determined by shared characteristics, as indicated through the comparison of genetic and anatomical similarities.

Phylogeny

- Evolutionary relationship among species.
- Represented as phylogenetic tree.



Types of Phylogeny

- **Phylogeny** can be of two types:
    1. Phenotypic phylogeny or morphological phylogeny
    2. Molecular phylogeny

Phenotypic Phylogeny

- It is also known as morphological phylogeny.
- Considered to be the traditional method.
- Based upon the phenotypic observation from the group of organism.
- Inefficient method to classify the micro-organism because phenotypic dissimilarity may be superficial.

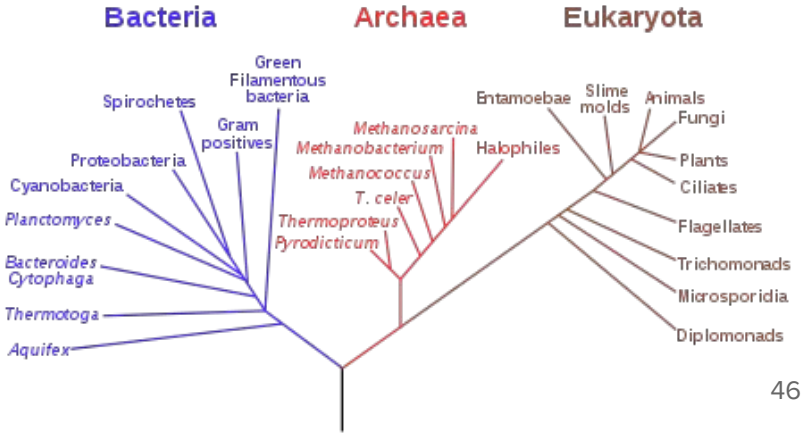
Molecular Phylogeny

- In **molecular phylogeny**, analysis of DNA and protein structure is used to determine genetic relationships among different organisms. For example, the analysis of cytochrome C, a protein in cell mitochondria that functions in the electron transport system and energy production, is used to determine degrees of relationship among organisms based on similarities of amino acid sequences in cytochrome C. Similarities in characteristics of biochemical structures, such as DNA and proteins, are then used to develop a phylogenetic tree based on inherited shared traits.
-

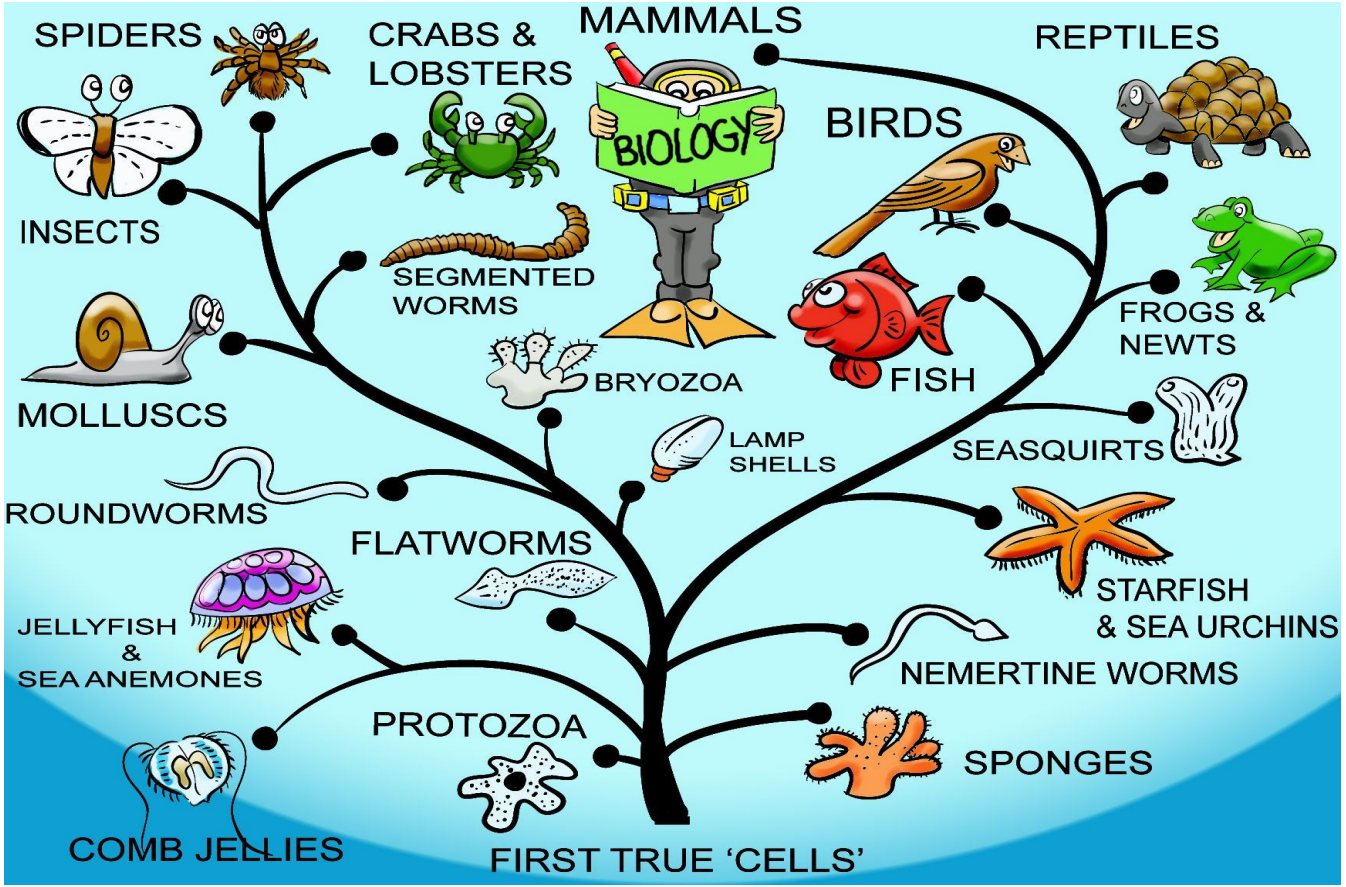
Phylogenetic Tree

- A phylogenetic tree (aka cladogram) is a diagrammatic representation of the evolutionary relatedness between various organisms, or at least our hypothesis regarding such. It is an important tool in understanding life as a whole and how various traits came about in living things. Phylogenetic tree is also known as **Denogram** or **tree of life.**
- Of course, we can't travel back in time to witness exactly how species diverged over time. However, we can rely upon an important clue present in all organisms today that reveals much about their past and the evolution of organisms, namely, the genome.

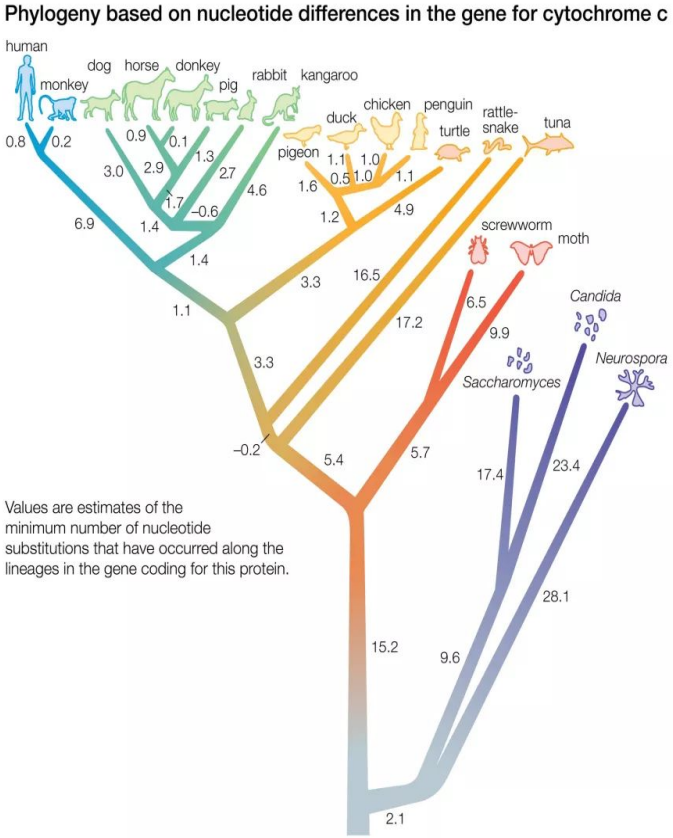
Life as we know it today is very diverse, ranging from bacteria to monkeys to insects, but one commonality between all living things is the presence of the genetic code. The differences between various organisms are reflected in their genomes; two more closely related organisms have a more related genome than two more distantly related organisms. Therefore, we can reason that if the genomes of species A and B are very similar, they have a recent common ancestor.



Phylogenetic Tree



Phylogenetic Tree

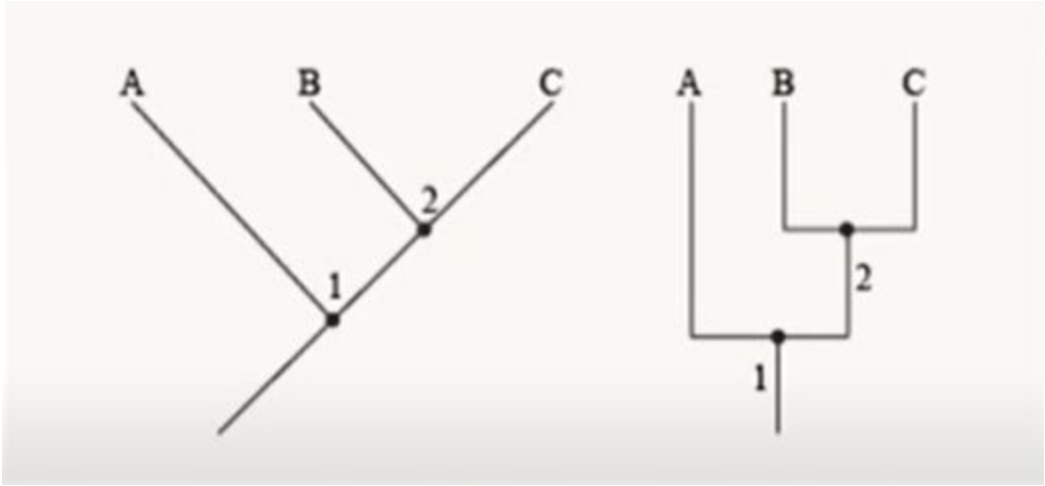


Phylogenetic Tree

Category	Genetic Similarity
Human and Human	99.9%
Human and Chimp	98.8%
Human and Pig	98%
Human and Mouse	97%
Human and Dogs	94%
Human and Cats	90%
Human and Cows	80%
Human and Fruit Flies	60%

Phylogenetic Tree

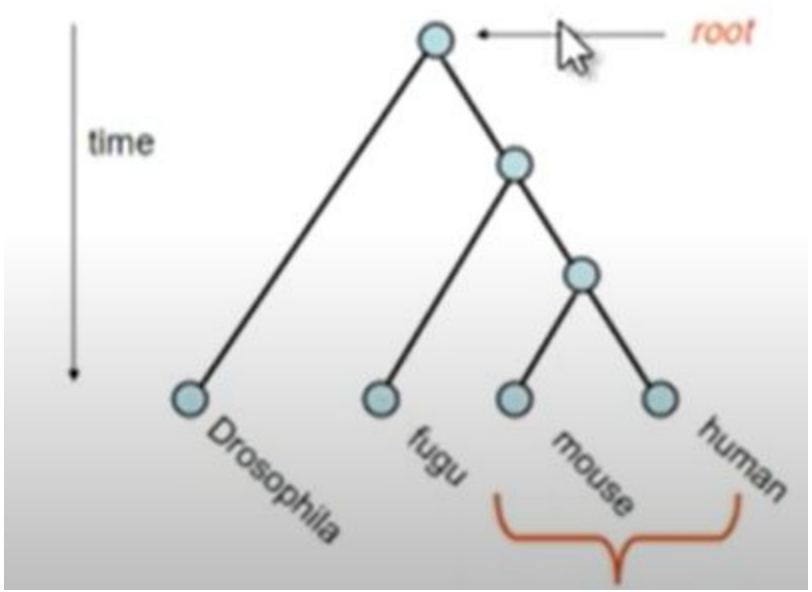
- **Phylogenetic trees can be drawn in two formats:**
  1. **Slanted format (angled form)**
  2. **Vertical format (squared form)**



Phylogenetic Tree

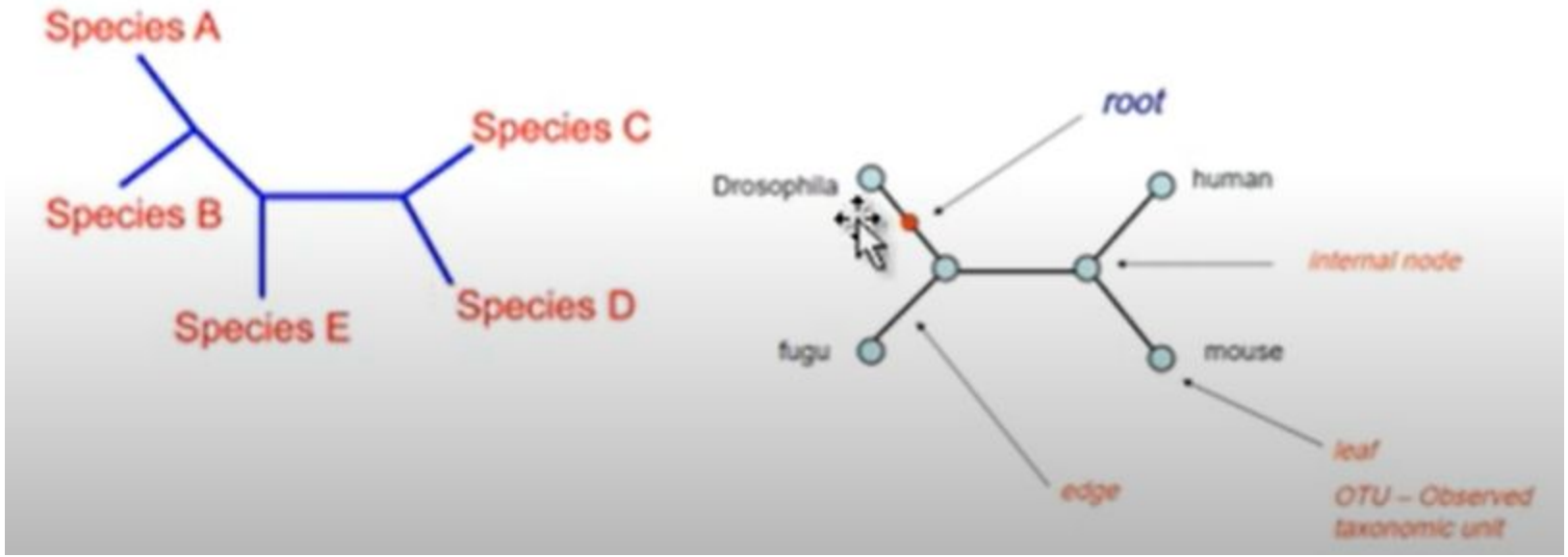
- **Rooted tree: having starting node or root.**
- **Root: common ancestor of all taxa.**

Clade: group of two or more taxa that include both their common ancestor and all of their descendants.



Phylogenetic Tree

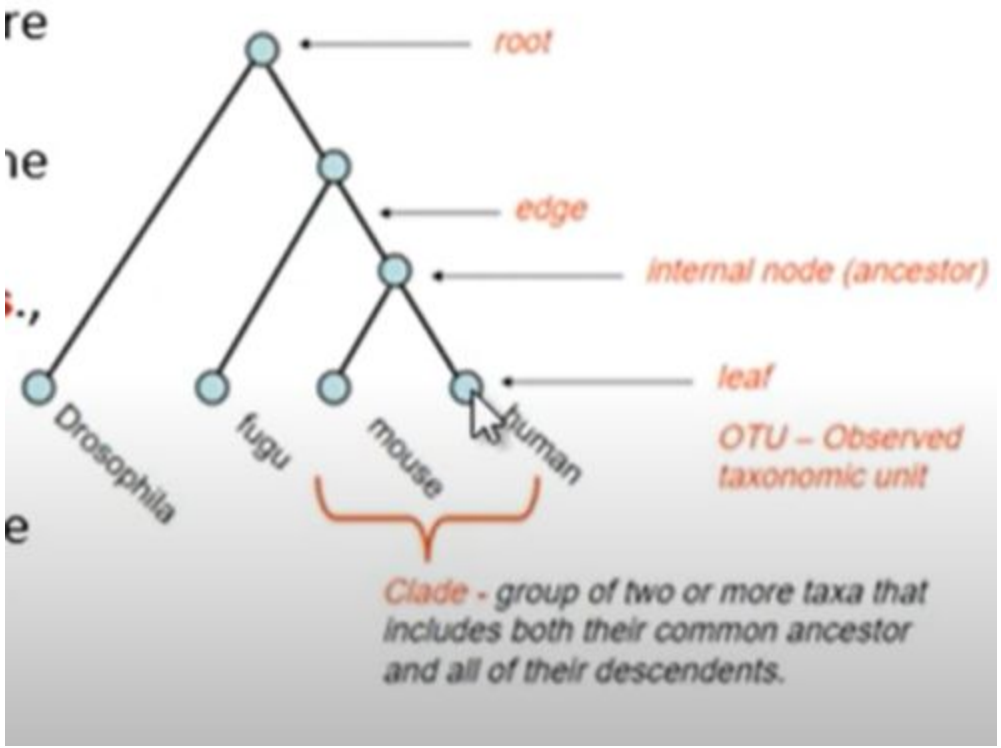
- **Unrooted tree: When no assumption is made about common ancestry, the corresponding phylogenetic tree does not have a unique root. Such a tree is called unrooted phylogenetic tree.**
-



Phylogenetic Tree

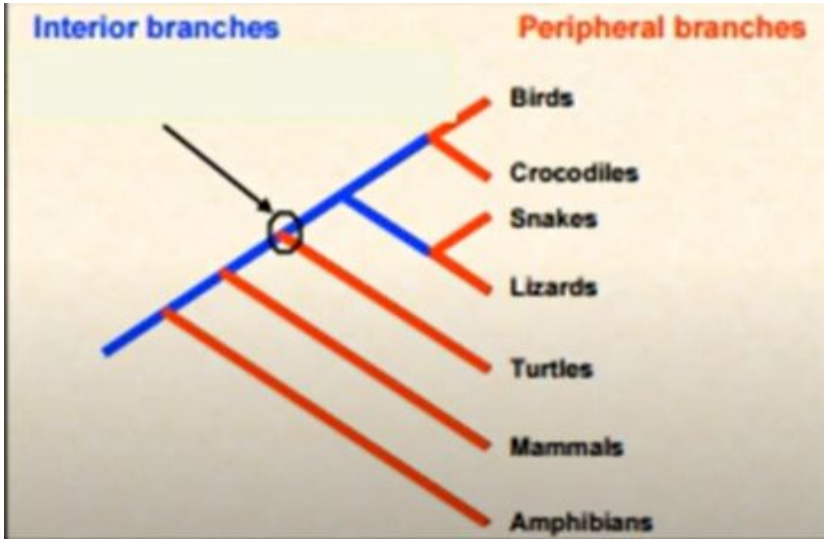
The lines in the tree are called branches/edges.

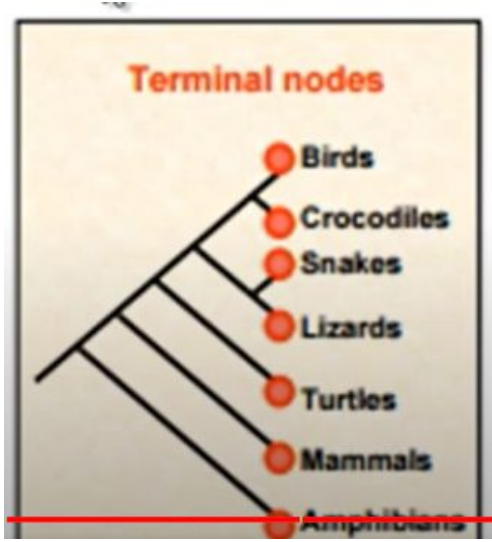
At the tip of the branches are present day species or sequences known as taxa (the singular form is taxon) or observed taxonomic units.
The bifurcating point at the very bottom of the tree is the root node, which represents the common ancestor of all members of the tree.

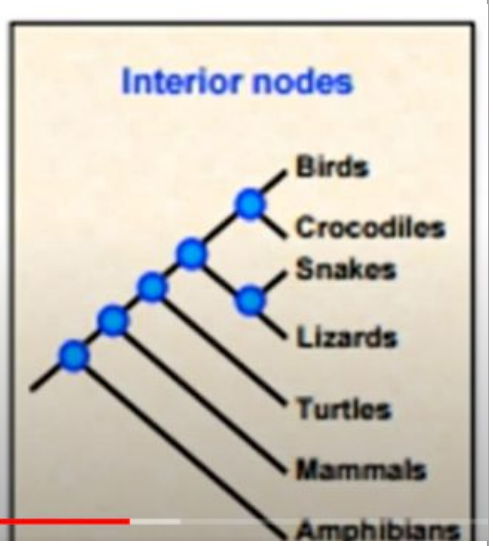


Phylogenetic Tree

The connecting point where two adjacent branches join is called a node.

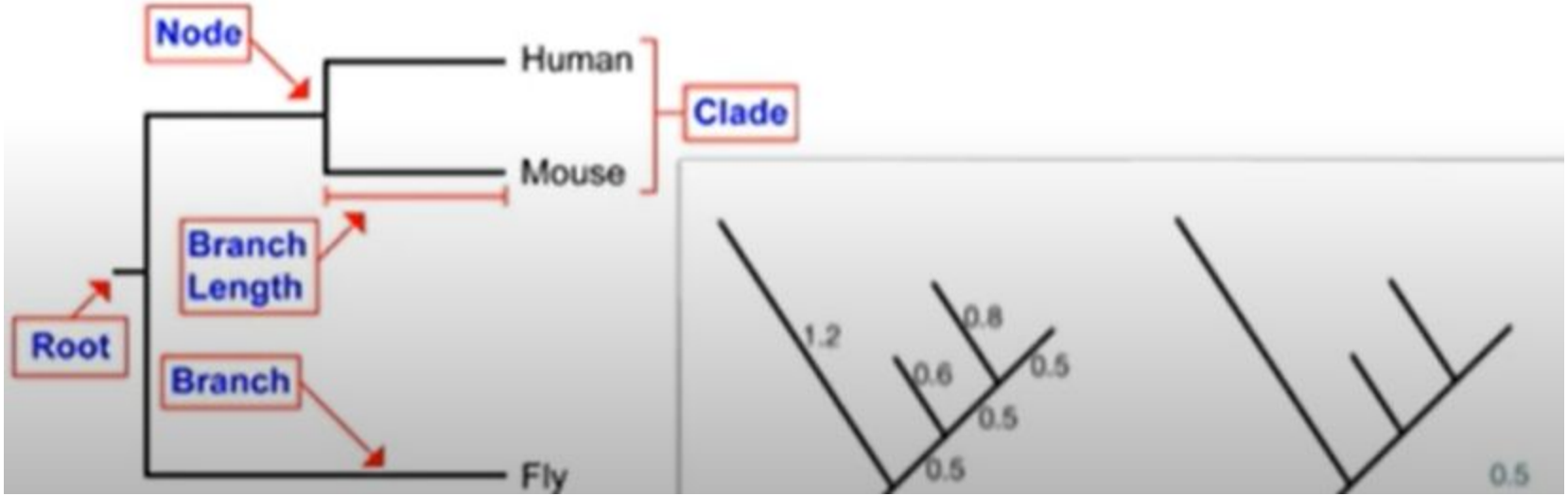






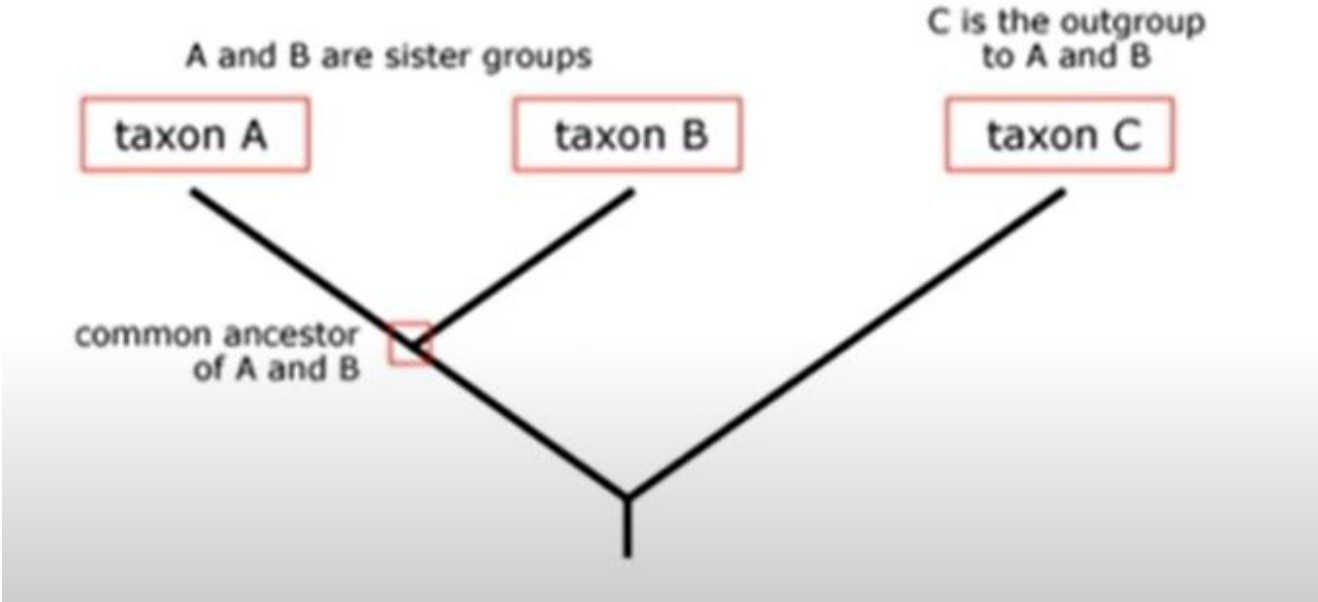
Phylogenetic Tree

Branch Length: The lengths of the branches of the tree represent evolutionary distances. Distances may be in units of time (million of years) but for sequence data, it is more common to see distances in unit of substitutions per site.



Phylogenetic Tree

Sister Group vs. Out Group

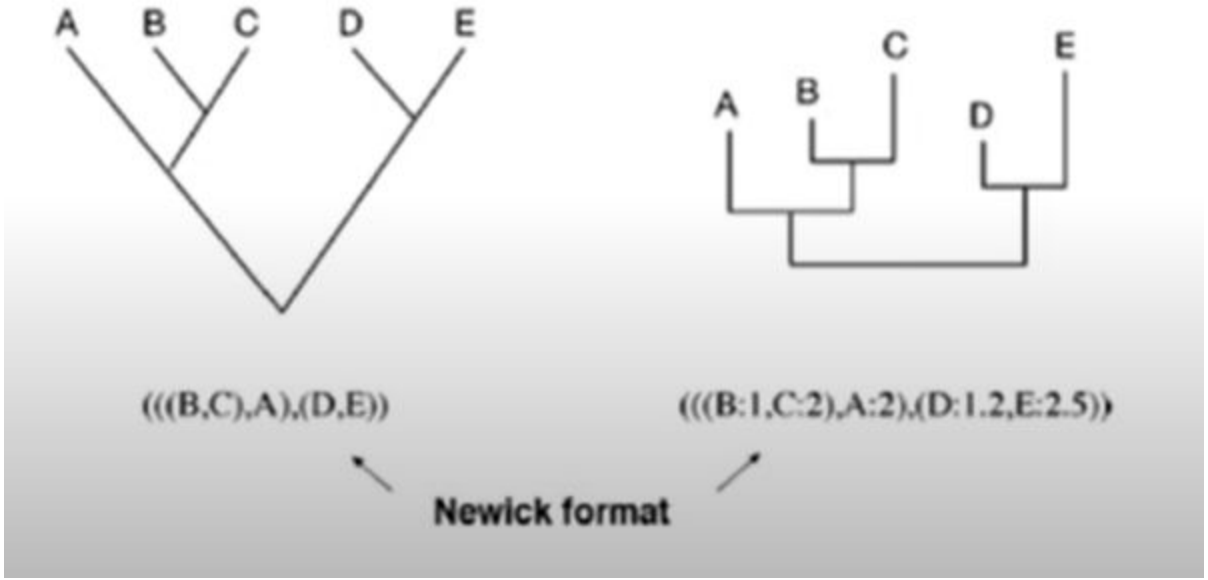


Phylogenetic Tree

The choice of which species goes to the right and which one goes to the left is arbitrary. Thus these phylogenetic trees are equivalent.

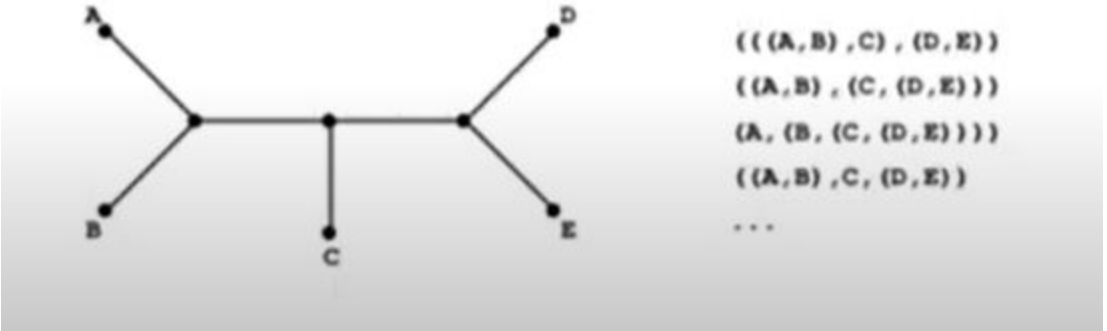
Phylogenetic Tree

Electronically, trees are usually held in a readable text file in the Newick format.



Phylogenetic Tree

There are many ways to represent unrooted trees.



UPGMA

Problem: Given the differences between a set of organisms or simply given a set of sequences of organisms, we have to create a phylogenetic tree.

Solution: There are many ways to create a phylogenetic tree, one of which is UPGMA (Unweighted Pair Group Method with Arithmetic Mean). This algorithm is very simple and can be boiled down to three simple steps:

1. Find the two organisms with least differences.
2. Group them together as one cluster and recalculate the differences.
3. Repeat step 1-2 until the tree is complete.

UPGMA implicitly assumes a constant substitution rate over time and phylogenetic lineages (known as molecular clock hypothesis). Since the assumption is often violated, this method is rarely used.

UPGMA

UPGMA

Step-0: Start aligning the sequences in pairwise fashion and note the difference between any pair of sequences in matrix.

	A	B	C	D	E	F
A		9	2	4	9	10
B			9	6	2	10
C				5	9	10
D					6	10
E						10
F						

Fig. Distance matrix represents the pairwise evolutionary distance among species

Fig. Given input (DNA) sequences of six different species.

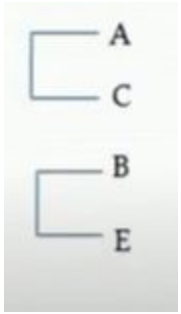
UPGMA: Iteration-I

Step-1: Find two sequences with least difference.

	A	B	C	D	E	F
A		9	2	4	9	10
B			9	6	2	10
C				5	9	10
D					6	10
E						10
F						

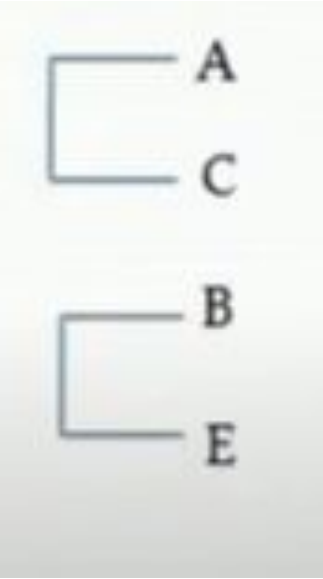
Fig. Distance matrix represents the pairwise evolutionary distance among species

Here, smallest distant pair of sequences are A-C and B-E. Hence, we group A and C into one cluster, and group B and E into another cluster and then update the distance matrix.



UPGMA: Iteration-I

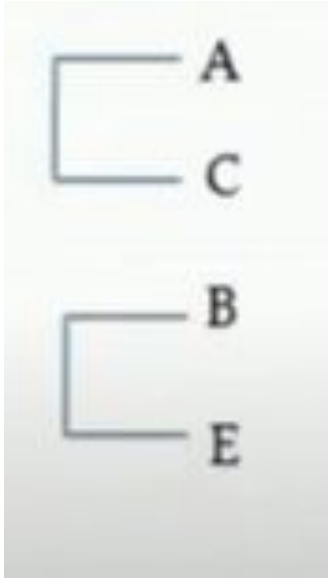
Now we update the distance matrix by first treating A-C as one composite sequence and calculate the mean distance of other sequences from it.



	A/C	B	D	E	F
A/C		9	4.5	9	10
B			6	2	10
D				6	10
E					10
F					

UPGMA: Iteration-I

Now, the distance matrix is again updated treating B and E as one composite sequence.

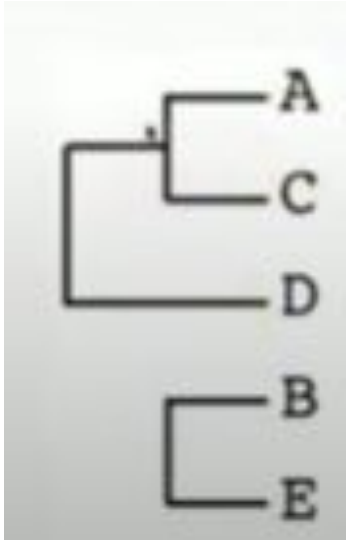


	A/C	B/E	D	F
A/C		9	4.5	10
B/E			6	10
D				10
F				

Now repeat step 1 and 2 until the tree is complete.

UPGMA: Iteration-II

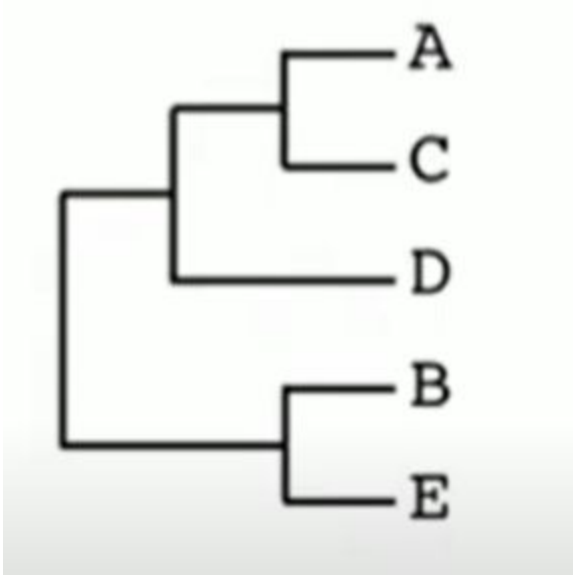
Now, the most closely related sequences are A-C and D. So we group them together to form a single cluster. And update the distance matrix treating A-C and D as one composite sequence (A/C/D)



	A/C/D	B/E	F
A/C/D		7.5	10
B/E			10
F			

UPGMA: Iteration-III

Now, the most closely related sequences are A/C/D and B/E. So, we group them together to form a single cluster and update the distance matrix treating A/C/D and B/E as one composite sequence (A/C/D/B/E).



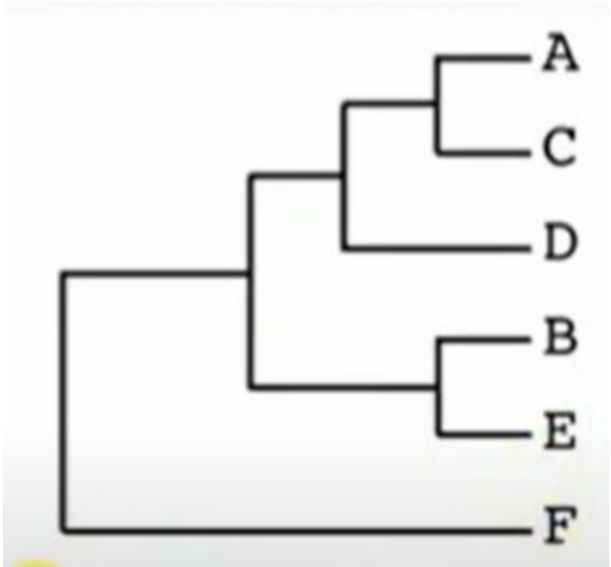


Fig. Final tree

UPGMA: Exercise

Problem: Given the distance matrix as shown below. Construct a phylogenetic tree to show the evolutionary relatedness between different species.

THE NUMBER OF AMINO ACID DIFFERENCES IN CYTOCHROME *c* AMONG VARIOUS ORGANISMS

	Horse	Donkey	Chicken	Penguin	Snake
Horse	0	1	11	13	21
Donkey		0	10	12	20
Chicken			0	3	18
Penguin				0	17
Snake					0

Neighbour Joining Method

- Neighbour Joining is a clustering approach for the reconstruction of the phylogenetic tree.
- Given by Saitou and Nei in 1987.
- Does not have the molecular clock theory assumption.

Key Idea:

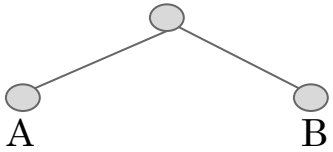
Find a pair of leaves that are close to each other but far from other leaves.

Neighbour Joining Method

UPGMA	Neighbour Joining
Divergence of sequences is assumed to occur at a constant rate.	Constructs an unrooted guide tree from a distance matrix.
Distance to root/ancestor is equal.	We don't assume constant rate of evolution. Hence, distance to root/ancestor might vary.

Neighbour Joining Method

- One of the important concepts in the Neighbour Joining method is neighbours, which are defined as two taxa connected by single node in an unrooted tree.



- **Input:**

  The input is the ***n*** number of taxa.

- **Output:**

  The output is an unrooted tree with branch lengths.
- It repeatedly joins pair of leaves (or subtrees) by rules of numerical optimization.
- It shrinks the distance matrix by considering two neighbours as one node.

Neighbour Joining Method

- Algorithm:
  1. For each terminal node ***i,*** compute:

  2.

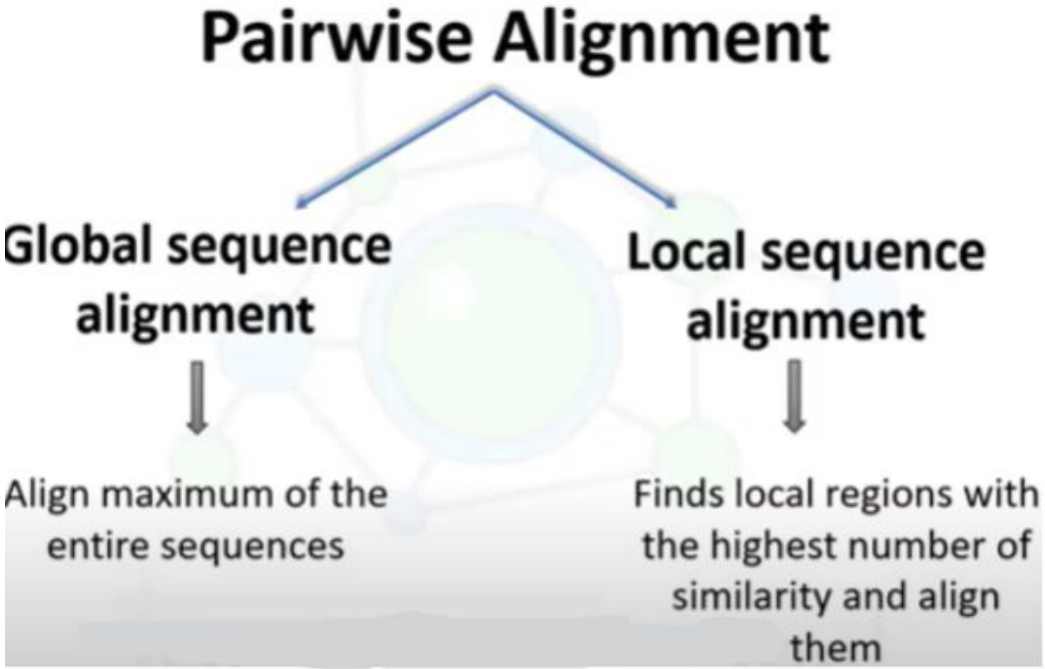
Sequence Alignment

- **Terms Used in Sequence Alignment/ Comparison:**

(Qry)	A C D E F G	A C D E F G	A C D E F G	A C -- E F G
(Sbj)	A C D E F G	A C L E F G	A C -- E F G	A C D E F G
Biological event	Conservation	Substitution	Insertion	Deletion
Alignment represent	Match	Mismatch	Gap	Gap

Sequence Alignment

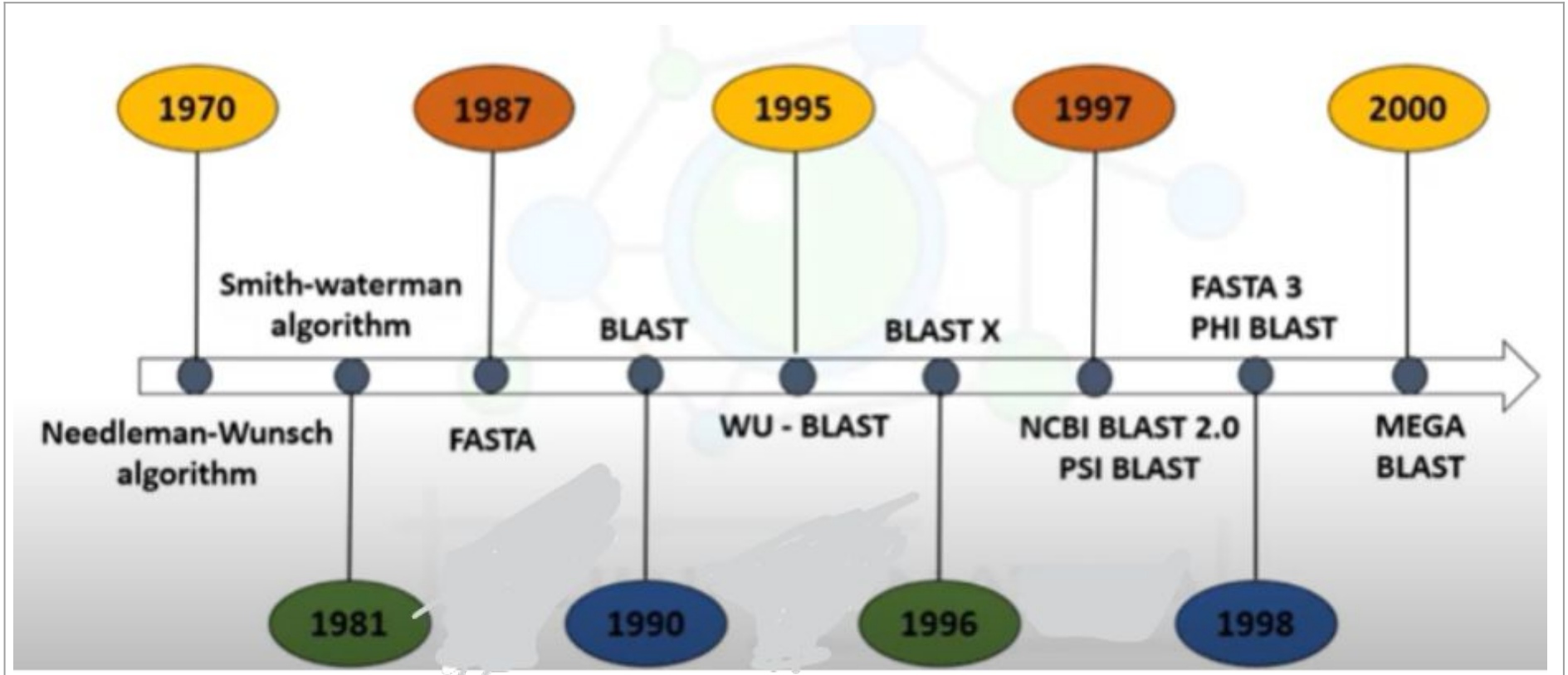
- Sequence alignment can be of two types:
  - **pairwise sequence alignment:** A pair of sequences (two sequences) are aligned.
  - **Multiple sequence alignment:** More than two sequences are aligned.



Sequence Alignment (Global and Local)

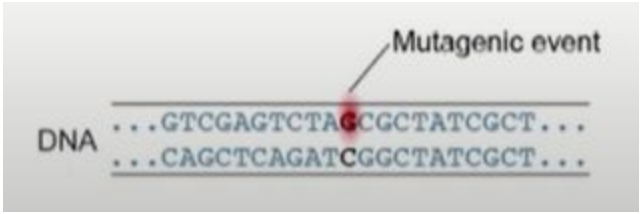
Global Alignment	Local Alignment
	
Tries to align entire sequence	Align regions with highest similarities.
Align all letters from query and target.	Align substring of target with substring of query.
Suitable for closely related sequences.	Suitable for divergent sequences.
A general global alignment method is **Needleman-Wunsch.**	A general local alignment method is **Smith-Waterman.**

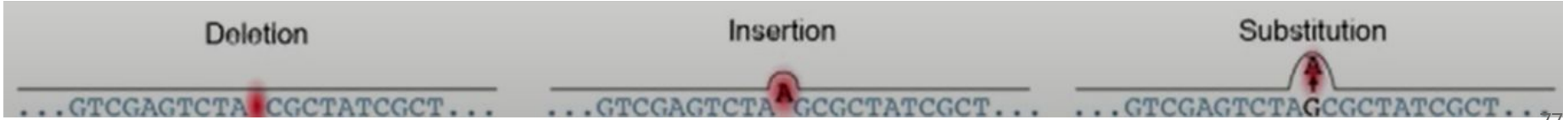
History of Alignment Algorithms



Biological Significance of Gaps

- The biological process of protein synthesis namely, replication, transcription and translation can produce errors resulting in mutations in the final nucleic acid sequence.
- Changes that occur during mutation are categorized into 3 classes. Such as:
    1. **Substitution:** Scoring matrix (PAM, BLOSSUM) capture substitutions.
    2. **Insertion**
    3. **Deletion**
- **Insertion and deletion:** both are dealt with **gap penalty.**





Biological Significance of Gaps

- Performing optimal alignment between sequences often involves applying gaps that represent insertions and deletions.
- In alignments, gaps are represented as contiguous dashes (-) on a protein/ DNA sequence alignment.
- Because in natural evolutionary process, insertion and deletion are relatively rare in comparison to substitutions, including gaps should be made more difficult computationally, reflecting the rarity of insertional and deletional events in evolution.
- Gaps are penalized via various gap penalty scoring methods.

- **What is Gap penalty?**
  It is the cost to introduce gap.

Biological Significance of Gaps

Gap Penalty

Gaps are not in nature. Gap is what we do in order to understand evolutionary relationships between sequences.

In order to incorporate gaps, we have gap penalty for each gap in the sequence.

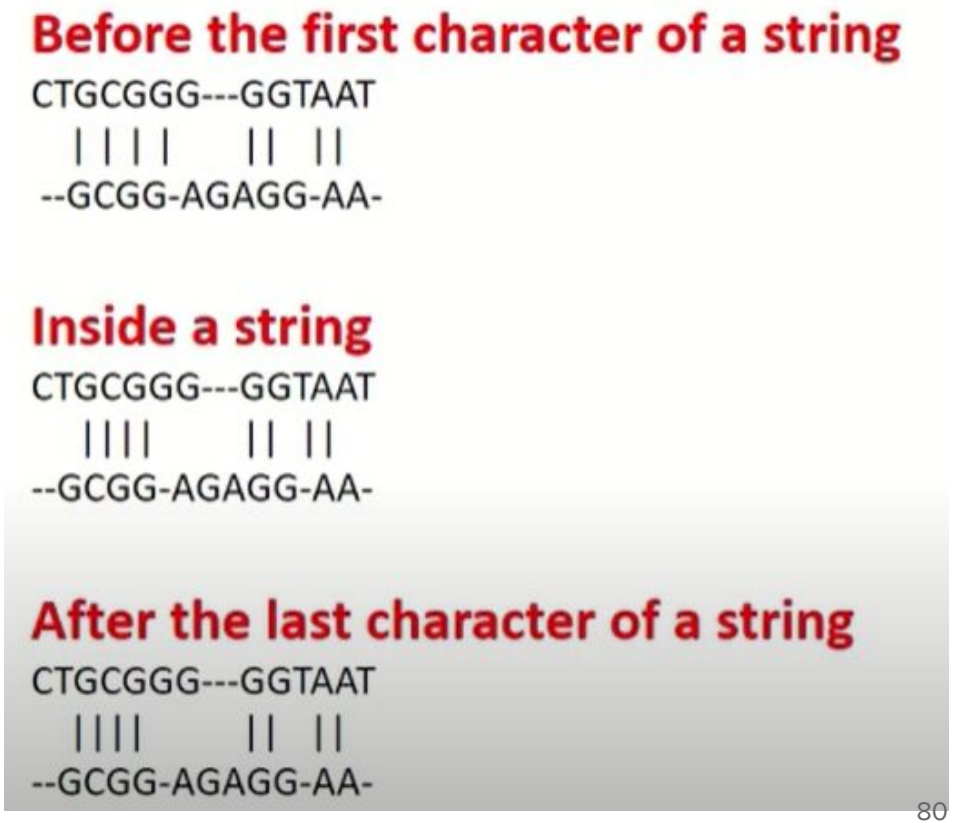
We must be very careful in adding gaps in the sequence. It would be good to have 1 mismatch than to have many gaps.

Biological Significance of Gaps

Gaps can occur at various different places in a sequence as depicted to the diagram right.

However, assigning penalty values can be more or less arbitrary because there is no evolutionary theory to determine a precise cost for introducing insertions and deletions.

If the penalty values are set too low, gaps can become too numerous to allow even unrelated sequences to be matched up with high similarity scores.



Biological Significance of Gaps

- If the penalty values are set too high, gaps can become too difficult to appear, and reasonable alignment cannot be achieved, which is also unrealistic.

When the gap penalty is high, fewer gaps will be inserted.	If you are searching for sequences that are a strict match for your query sequence, the gap penalty should be set high. This will retrieve regions with very closely related sequences.
When the gap penalty is low, more and larger gaps will be inserted.	If you are searching for similarity between distantly related sequences, the gap penalty should be set low.

Biological Significance of Gaps

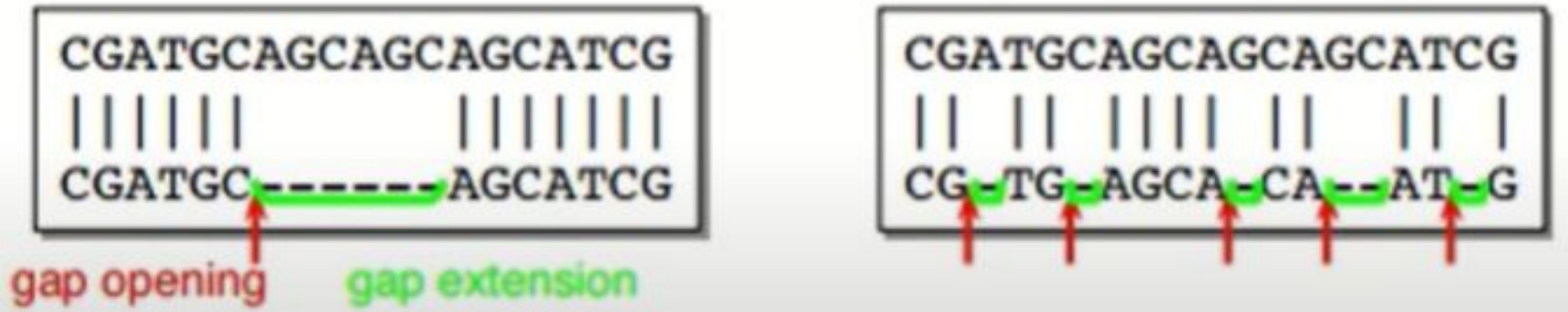
- Through empirical studies for globular proteins, a set of penalty values have been developed that appear to suit most alignment process.
- They are normally implemented as default values in most alignment programs.
- Another factor to consider is the cost difference between opening a gap and extending a gap.

- **Gap opening penalty:**
  Counted each time a gap is opened in an alignment.
- **Gap extension penalty:**
  Counted for each extension of a gap in an alignment.

- It is known that it is easier to extend a gap that has already been started. Thus, gap opening should have a much higher penalty than gap extension.

Biological Significance of Gaps

- Two alignments with identical number of gaps but very different gap distribution. We may prefer one large gap over several small ones. (e.g. poorly conserved loops between well-conserved helices)

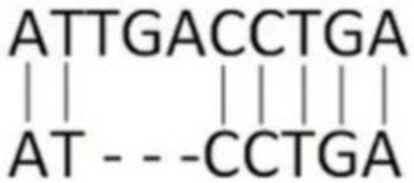


Types of Gap Penalty

- The three main types of gap penalties are: **constant**, **linear** and **affine gap penalty**.

- **<u>Constant Gap Penalty:</u>**
  - **This is the simplest type of gap penalty.**
  - **Assign the same negative score for each gap position regardless whether it is opening or extending.'**
  - **However, this penalty scheme has been found to be less realistic.**

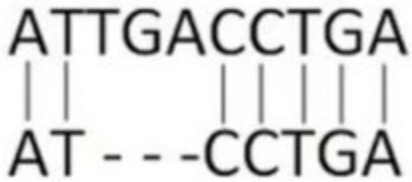
Aligning two short DNA sequences, with "-" depicting a gap of one base pair. If each match was worth 1 point and the gap -1, the total score: 7-1 = 6



Types of Gap Penalty

- **Linear Gap Penalty:**
  - **The linear gap penalty takes into account the length (L) of each insertion/deletion in the gap.**
  - **Therefore, if the penalty for each inserted/deleted element is B and the length of the gap is L, then total gap penalty = B*L.**
  - **However, this penalty scheme has been found to be less realistic.**

Unlike constant gap penalty, the size of the gap is considered.
With a match with score 1 and gap -1, the total score: 7-3 = 64



Types of Gap Penalty

- **Affine Gap Penalty:**
  - **The most widely used gap penalty function is the affine gap penalty.**
  - **The affine gap penalty combines the components in both the constant and linear gap penalty. For example, one may use a -12/-1 scheme in which the gap opening penalty is -12 and the gap extension penalty is -1.**

$$W = \gamma + \delta \times (k - 1)$$

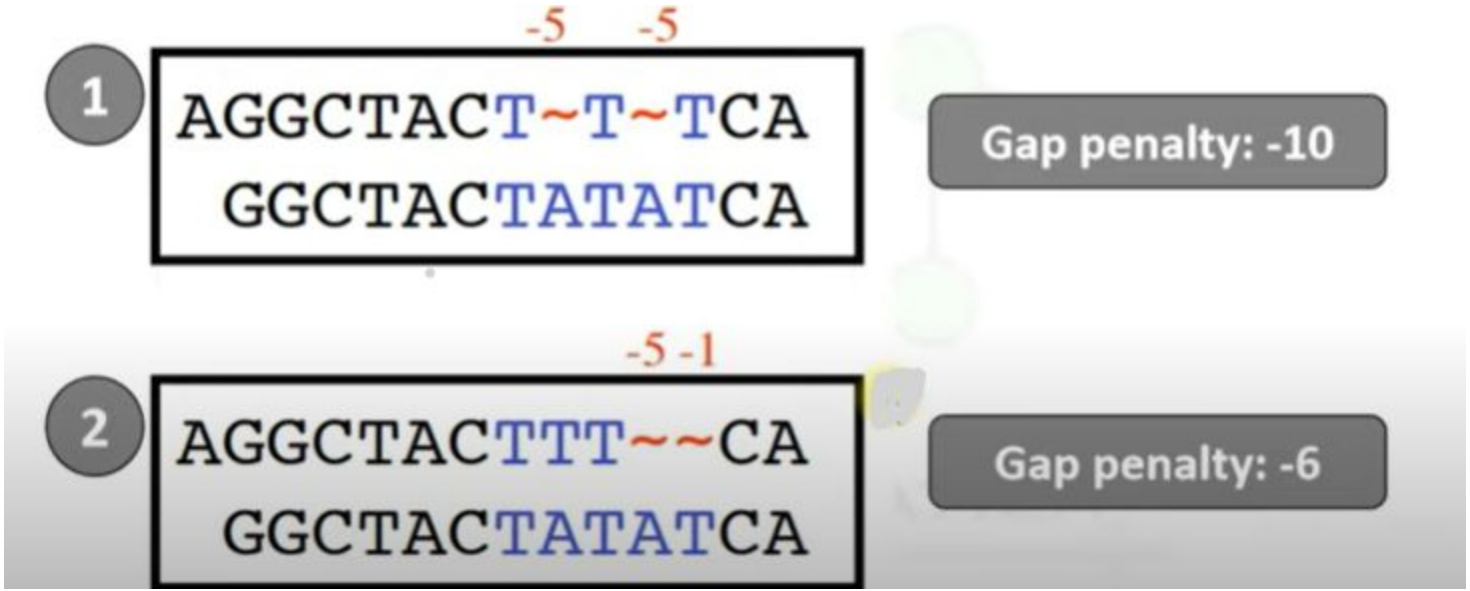
  - **The total gap penalty (W) is a linear function of gap length which is calculated as:**
  - **Where** γ is the gap opening penalty, δ is the gap extension penalty, and k is the length of the gap.

Types of Gap Penalty

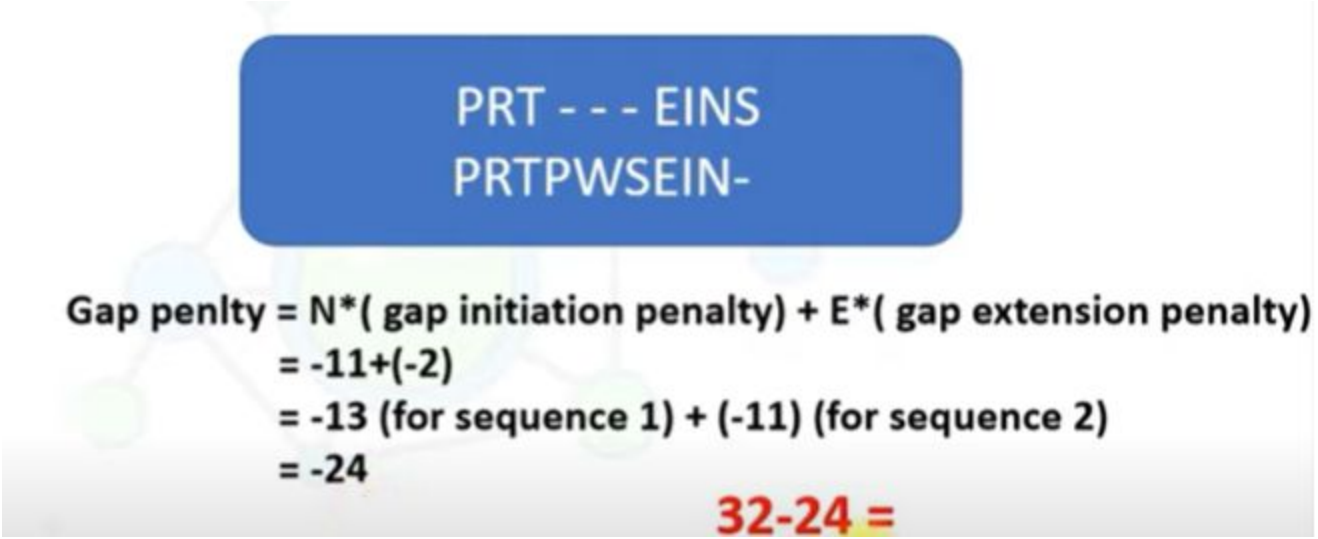
```
G A A T T C C G T T A
|   | |   |   |     |
G G A T _ C _ G _ _ A
              v
+ - + + - + - + - - +
2 1 2 2 2 2 2 2 2 1 2   = score 4
```

```
G A A T T C C G T T A
|   | |   > | |     |
G G A T _ _ C G _ _ A
      v       v
+ - + + - - + + - - +
2 1 2 2 2 1 2 2 2 1 2   = score 5
```

Affine Gap Penalty



Affine Gap Penalty



Affine Gap Penalty

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

PRT - - - EINS

PRTPWSEIN-

32-24 =

Why?

Gaps at the Terminal Regions

- Gaps at the terminal regions are often treated with no penalty because in reality many true homologous sequences are of different lengths.
- Consequently, end gaps can be allowed to be free to avoid getting unrealistic alignments.

Gaps at the Terminal Regions

- Gaps at the terminal regions are often treated with no penalty because in reality many true homologous sequences are of different lengths.
- Consequently, end gaps can be allowed to be free to avoid getting unrealistic alignments.

Scoring Alignment

Why Scoring Alignment!

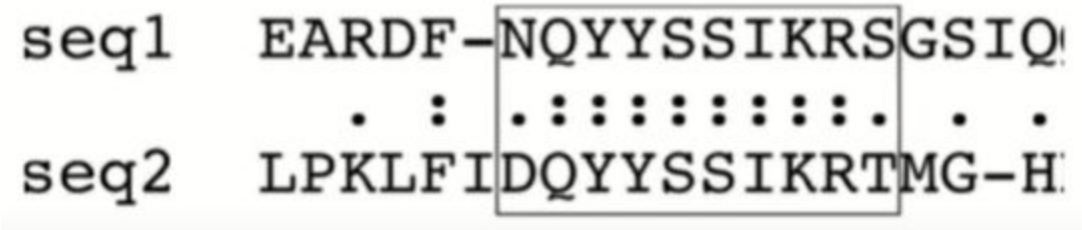
- **We need to differentiate good alignments from poor ones.**
- **We use a rule that assigns a numerical score to any alignment; the higher the score, the better the alignment.**
- **Sequence Similarity and Identity**
- **Sequence Alignment Score**

Sequence Similarity and Identity

- After an alignment is made, we can extract two quantitative parameters from each pairwise comparison:
  - Sequence similarity
  - Sequence identity
- Sequence similarity and sequence identity are synonymous for nucleotide sequences.
- For protein sequences, however, the two concepts are very different.

Sequence Identity

- In a protein sequence alignment, sequence identity refers to the percentage of matches of the same amino acid residues between two aligned sequences.

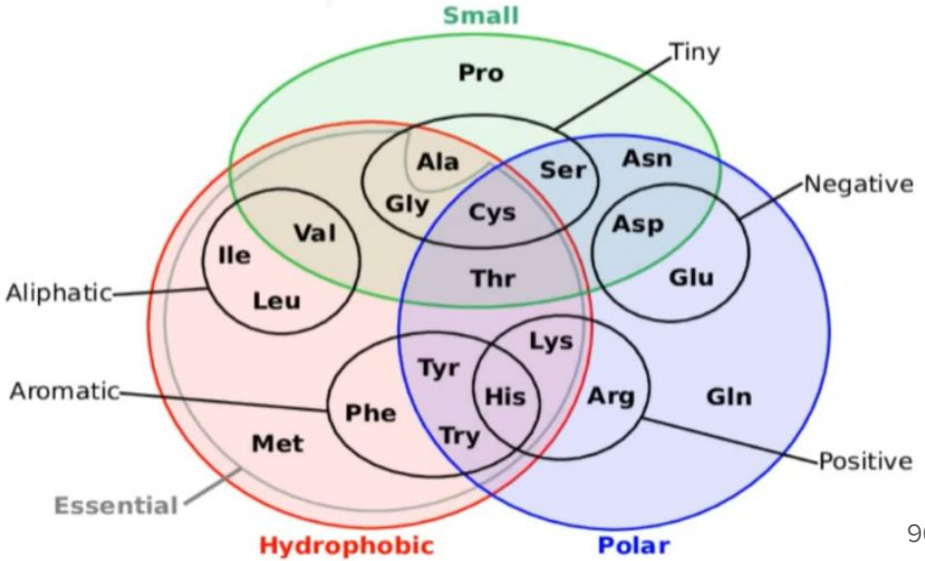


- In the line between the two sequences, ":" indicates identical residue matches and "." indicates similar residue matches.

Sequence Similarity

- **Similarity refers to the percentage of aligned residues that have similar physicochemical characteristics and can be more readily substituted for each other.**
- **When one amino acid is mutated to a similar residue such that the physicochemical properties are preserved, a conservative substitution is said to have occurred.**

For example, a change from arginine to lysine maintains the +1 positive charge. This is far more likely to be acceptable since the two residues are similar in property and won't compromise the translated protein. Thus the percentage similarity of two sequences is the sum of both identical and similar matches (residues that have undergone conservative substitution).



Calculation of the Sequence Similarity/Identity

- There are two ways to calculate the sequence similarity/identity:
  - One involves the use of the overall sequence lengths of both sequences.
  - The other normalizes by the size of the shorter sequence.

First Method

- **The first method uses the below formula:**

$$S = [(L_s \times 2)/(L_a + L_b)] \times 100$$

- **Where S is the percentage sequence similarity, Ls is the number of aligned residues with similar characteristics, L_a and L_b are the total lengths of each individual sequence.**

- **The sequence identity can be calculated in a similar fashion:**

$$I = [(L_i \times 2)/(L_a + L_b)] \times 100$$

First Method

Example:

The alignment is 39 amino acids long, and the human & fruit fly sequences differ at 1 position.

Human and fruit fly sequences have a percent identity of

$$= \frac{\text{Number of identical amino acids}}{\text{length of the alignment}} \times 100$$

$$= (38*100/39) = 97\%$$

First Method

Calculation of percent Similarity

A W G H E

A W - H E

% Similarity = (Number of matches / No of amino acids in region) * 100

% Similarity = 4/5 x 100

= 80%

Second Method

- **The second method of calculation is to derive the percentage of identical/similar residues over the full length of the smaller sequence using the formula:**

$$I(S)\% = L_{i(s)} / L_a \%$$

- **Where L_a is the length of the shorter of the two sequences.**

Second Method

- **Example: Say, sequence A has 320 AA, while sequence B has 450 AA. Using BLAST to perform a pairwise alignment, we see that 100 amino acids are identical. Thus % identity is**

$$\text{Identity} = 100 / 320 = 31.25\%.$$

- **Additionally, we see that there are 23 amino acids that are different by conservation substitution, meaning that their chemical properties are maintained. Hence, similarity is calculated as:**

$$\text{Similarity} = (100 + 23) / 320 = 38.44\%$$

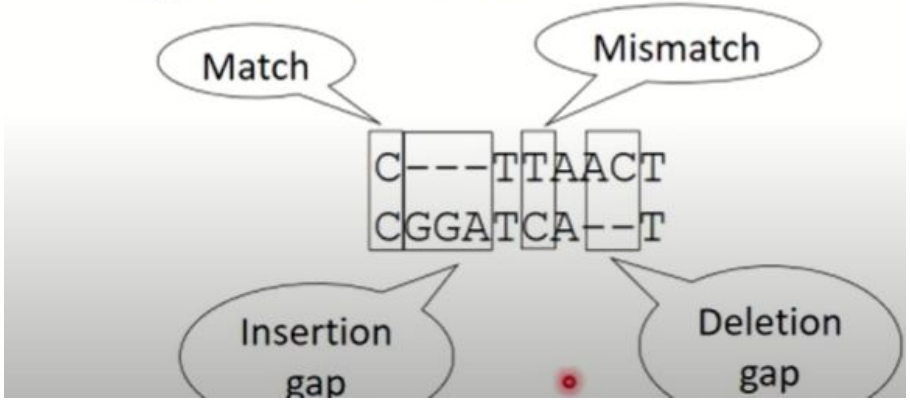
- **Thus, our sequence are 31.25% identical and 38.44% similar. Similarity is always greater than identity.**

Sequence Alignment Score

- **Sequence similarity score is calculated based on the pairwise alignment quality.**
- **Alignment score is the sum of scores for each position.**

Sequence A: CTTAACT
Sequence B: CGGATCAT

An alignment of A and B:



Sequence Alignment Score (Example-1)

```
AACGTTTCTCCAAATAGCTAGGC
===--=====-===-==-======
AACCGTTCTACAATTACCTAGGC
```

Hits(+1): 18

Misses (-2): 5

Score = 18 * 1 + 5 * (-2) = 8

Sequence Alignment Score (Example-2)

- Match: +8
- Mismatch: -5
- Each gap symbol: -3

C	–	–	–	T	T	A	A	C	T	
C	G	G	A	T	C	A	–	–	T	
+8	-3	-3	-3	+8	-5	+8	-3	-3	+8	= +12

Sequence Alignment Score (Example-3)

Alignment of ATCGGATCT and ACGGACT

```
A T C G G A T C T
|   | | |     | |
A - C G G - A C T
```

match: +2
mismatch: -1
indel –2

6 matches, 1 mismatch, and 2 indels

$6 * 2 + 1 * -1 + 2 * -2 = 7$

Sequence Alignment Score (Example-4)

- match: +4
- mismatch: -5
- gap opening: -16
- 4+4+(-5)+4+(-16)+4+4+4+4+4 = 11

```
A C G A C T G G C A
A C T A - T G G C A
```

Sequence Alignment Scoring Scheme

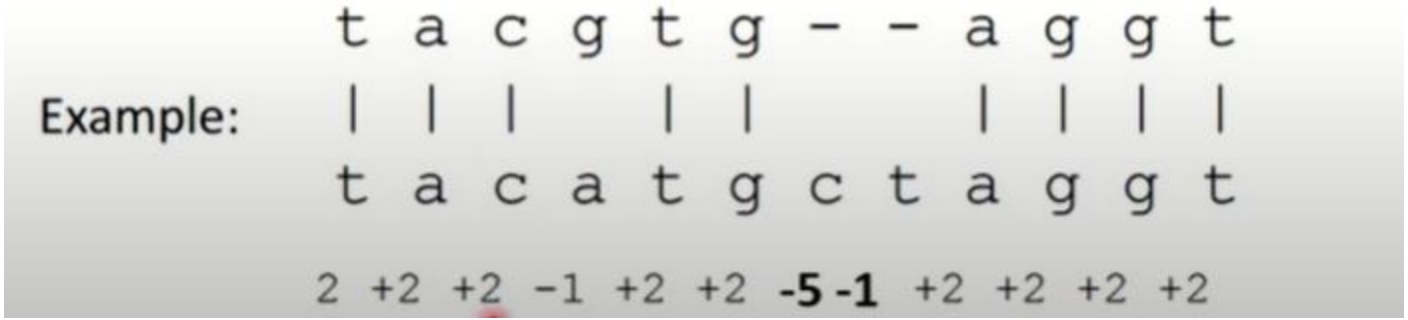
Substitution score matrix

	a	c	g	t
a	2	-3	-1	-3
c	-3	2	-3	-1
g	-1	-3	2	-3
t	-3	-1	-3	2

Gap scores

Gap existence cost: 5
Gap extension cost: 1

Alignment score = 11



Scoring Matrix

- We perform sequence alignment to determine similarity between sequences. One sequence may show similarity with many sequences and to determine the most similar sequence among them, we must provide a score to each similar sequence and this score is provided using scoring matrices. ‘
- Scoring matrix gives the degree of biological relationship between Amino acids or nucleotides.
- Each entry in the scoring matrix represents the probability that 2 AA occurs in homologous position in sequence that share a common ancestor
- Different AA partially match in chemical properties for which the scoring scheme of 1 for a match and 0 for a mismatch is not enough because Met→Leu substitution does not alter the hydrophobic interaction but Met→Arg alters the hydrophobic interaction as Arginine is charged. Hence, during the course of evolution, Met→Leu is more like to occur than Met→Arg. There, we should give more score to Met→Leu substitution.

AA Substitution Matrix (PAM & BLOSUM)

- Each matrix position is filled with a score.
- Score reflects how often on e AA would have been paired with the other in an alignment of related protein sequences.
- Probability of (A→B) = Probability of (B→A)
- Likelihood of replacement depends on:
    - The product of the frequency of occurrence of two.
    - Their chemical and physical properties.

PAM Matrix

Assumptions:

- Each change in the current AA at a particular site is assumed to be independent of the previous mutational events at that site.
- AA substitutions are viewed as a Markov model.
- A series of change of state in a system such that change from one state to another does not depend on the previous history of the state.
- AA substitutions observed over short periods of evolutionary history can be extrapolated to longer distances.

PAM1: A PAM unit is a time period over which 1% amino acids in a sequence are expected to undergo accepted mutations some of which may occur in the same position.

PAM1 means that the probability of changing one amino acid into another is 1% and the probability of not changing is 99%.

PAM Matrix (Substitution Scoring Matrix)

- A **point accepted mutation** — also known as a PAM — is the replacement of a single amino acid in the primary structure of a protein with another single amino acid, which is accepted by the processes of natural selection.
- A **PAM matrix** is a matrix where each column and row represents one of the twenty standard amino acids. In bioinformatics, PAM matrices are sometimes used as substitution matrices to score sequence alignments for proteins.
- The mutation data were accumulated from the phylogenetic trees and from a few pair of related sequences.
- If a substitution between two AA is observed frequently, then positions in which these two residues are aligned are scored favorably. Likewise, alignments between residues that are not observed to interchange frequently in natural evolution are penalized.
- Each entry in a PAM matrix indicates the likelihood of the amino acid of that row being replaced with the amino acid of that column through a series of one or more point accepted mutations during a specified evolutionary interval, rather than these two amino acids being aligned due to chance. Different PAM matrices correspond to different lengths of time in the evolution of the protein sequence.

Mutability of Amino Acids

- Relative mutability of amino acid is proportional to the ratio of changes to occurrence.

```
                    A G L L
                    A G A V

Amino acids:        A       G       L
Changes:            1       0       2
Frequency of occurrence:    3       2       2
Relative mutability:        0.33    0       1
```

The relative mutability of amino acids

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

Note that alanine is normalized to a value of 100.
Trp and cys are least mutable.
Asn and ser are most mutable.

Sample computation of relative mutability. The two aligned sequences may be two experimentally observed sequences or an observed sequence and its inferred ancestor.

PAM Matrix

PAM1 Mutation Matrix

1 PAM evolutionary distance

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

An element of this matrix, [Mij], gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 PAM. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

PAM1 is constructed using 71 group of related protein sequences, 85% similarity and 1572 amino acid changes.

PAM250 Matrix

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg	R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn	N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp	D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys	C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln	Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu	E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly	G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His	H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile	I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu	L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys	K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met	M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe	F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro	P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser	S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr	T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	33	6
Trp	W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr	Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val	V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

PAM 250

An element of this matrix, [Mij], gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 PAM. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

BLOSUM Matrix

- BLOSUM stands for Blocks Substitution Matrices.
- Derived from local, ungapped alignments of distantly related sequences.
- All matrices are calculated. No extrapolations are used.
- In bioinformatics, the BLOSUM matrix is a substitution matrix used for sequence alignment of proteins. BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences. They are based on local alignments.
- BLOSUM matrices were first introduced in a paper by Steven Henikoff and Jorja Henikoff. They scanned the BLOCKS database for very conserved regions of protein families (that do not have gaps in the sequence alignment) and then counted the relative frequencies of amino acids and their substitution probabilities.
- Then, they calculated a log-odds score for each of the 210 possible substitution pairs of the 20 standard amino acids. All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins like the PAM Matrices.

BLOSUM Matrix

Henikoff and Henikoff introduced BLOSUM matrix which led to marked improvements in alignments.

BLOSUM matrices with high numbers are designed for comparing closely related sequences, while those with low numbers are designed for comparing distantly related sequences. For example, BLOSUM 80 is used for closely related alignments, and BLOSUM 45 is used for more distantly related alignments.

BLOSUM r: the matrix is built from blocks with less than r% similarity. For example, BLOSUM 62 is the matrix built using sequences with less than 62% similarity (Sequences with similarity >= 62% were clustered.)

BLOSUM 62 is the default matrix for protein BLAST. Experimentation has shown that the BLOSUM62 matrix is among the best for detecting most weak protein similarities.

The BLOSUM matrices are derived from BLOCKS database. BLOCKS database is the set of ungapped alignment of sequence regions of related protein. The ***r*** represents the cutoff value for clusters the related protein sequences with greater than r% similarity. If one wants more diverse sequences to be included in the cluster, lower cut-off value must be selected, because lower cutoff value represents longer evolutionary timescales.

The Blocks Database

The Blocks Database contains multiple alignments of conserved regions in protein families.

Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.

The blocks for Blocks Database are made automatically by looking for the most highly conserved regions in groups of proteins represented in the PROSITE database. These blocks are then calibrated against the SWISS-PROT database to obtain a measure of the random distribution of matches. It is these calibrated blocks that make up the Blocks database.

The database can be searched to classify protein and nucleotide sequences.

BLOSUM62 Matrix

The BLOSUM62 matrix, the amino acids have been grouped and coloured based on Margaret Dayhoff's classification scheme. Positive and zero values have been highlighted.

	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
A	0	1	0	4																	A
G	-3	0	-2	0	6																G
P	-3	-1	-1	-1	-2	7															P
D	-3	0	-1	-2	-1	-1	6														D
E	-4	0	-1	-1	-2	-1	2	5													E
Q	-3	0	-1	-1	-2	-1	0	2	5												Q
N	-3	1	0	-2	0	-2	1	0	0	6											N
H	-3	-1	-2	-2	-2	-2	-1	0	0	1	8										H
R	-3	-1	-1	-1	-2	-2	-2	0	1	0	0	5									R
K	-3	0	-1	-1	-2	-1	-1	1	1	0	-1	2	5								K
M	-1	-1	-1	-1	-3	-2	-3	-2	0	-2	-2	-1	-1	5							M
I	-1	-2	-1	-1	-4	-3	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-1	-4	-3	-4	-3	-2	-3	-3	-2	-2	2	2	4					L
V	-1	-2	0	0	-3	-2	-3	-2	-2	-3	-3	-3	-2	1	3	1	4				V
W	-2	-3	-2	-3	-2	-4	-4	-3	-2	-4	-2	-3	-3	-1	-3	-2	-3	11			W
Y	-2	-2	-2	-2	-3	-3	-3	-2	-1	-2	2	-2	-2	-1	-1	-1	-1	2	7		Y
F	-2	-2	-2	-2	-3	-4	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	1	3	6	F
	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	

Relationship b/w PAM and BLOSUM

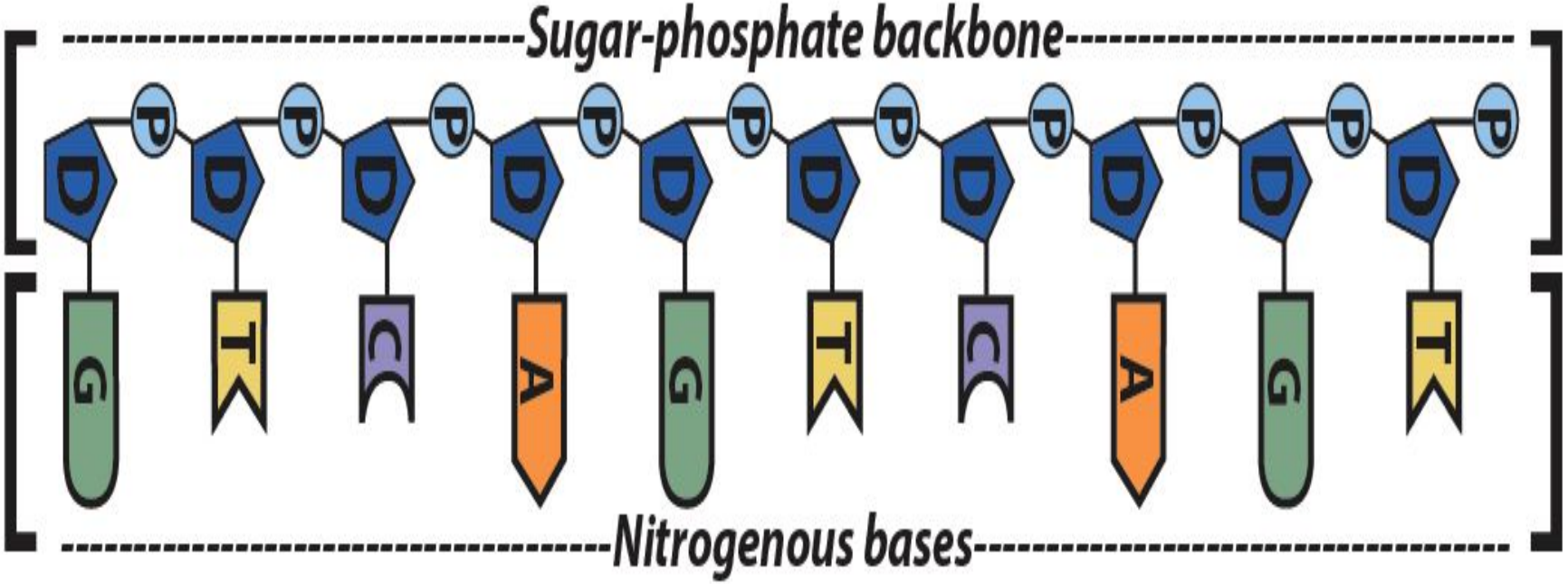
PAM	BLOSUM
To compare closely related sequences, PAM matrices with lower numbers are created.	To compare closely related sequences, BLOSUM matrices with higher numbers are created.
To compare distantly related proteins, PAM matrices with high numbers are created.	To compare distantly related proteins, BLOSUM matrices with low numbers are created.

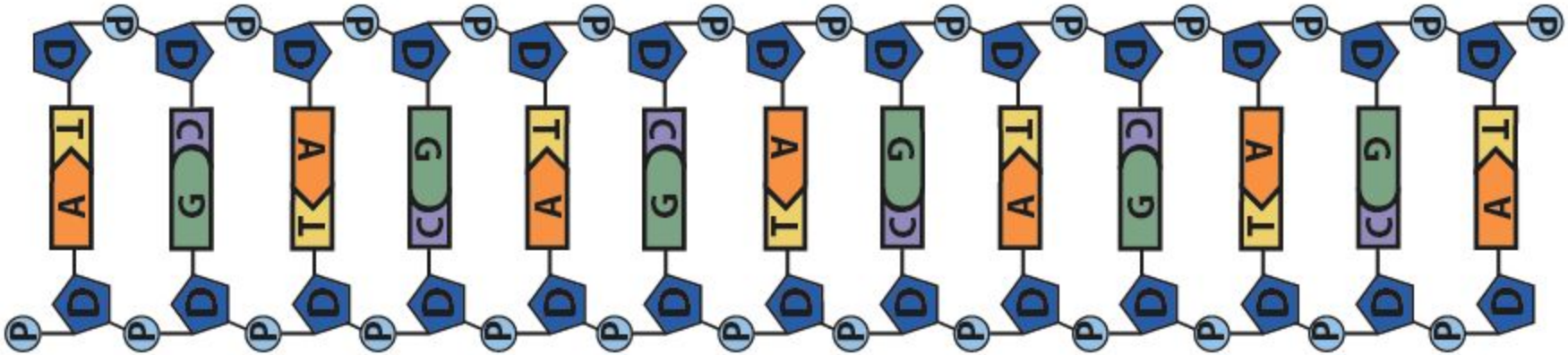
Difference b/w PAM and BLOSUM

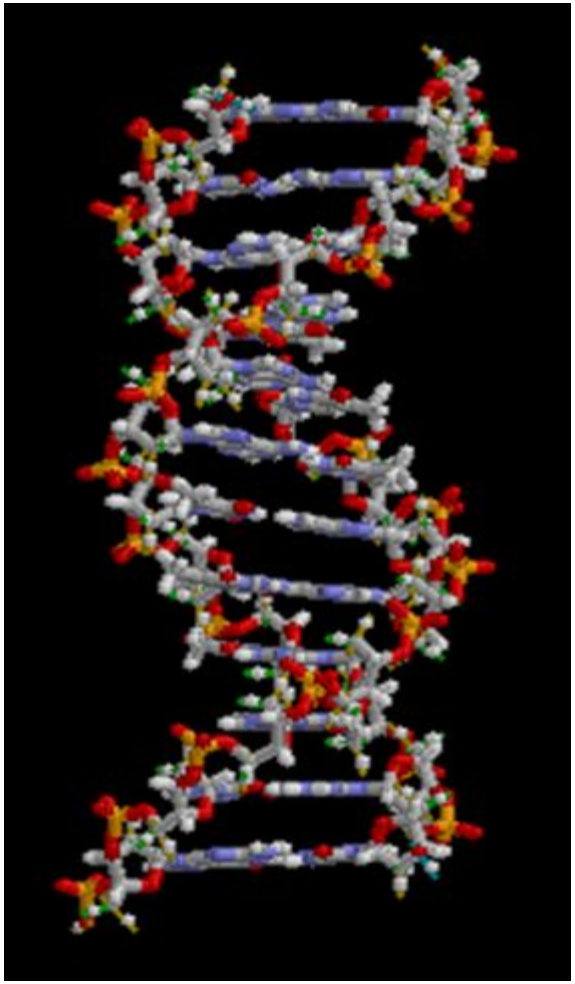
PAM	BLOSUM
Based on global alignments of closely related proteins.	Based on local alignments.
PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence but corresponds to 99% sequence identity.	BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.
Other PAM matrices are extrapolated from PAM1.	Based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
Higher numbers in matrices naming scheme denote larger evolutionary distance.	Larger numbers in matrices naming scheme denote higher sequence similarity and therefore smaller evolutionary distance.[19]

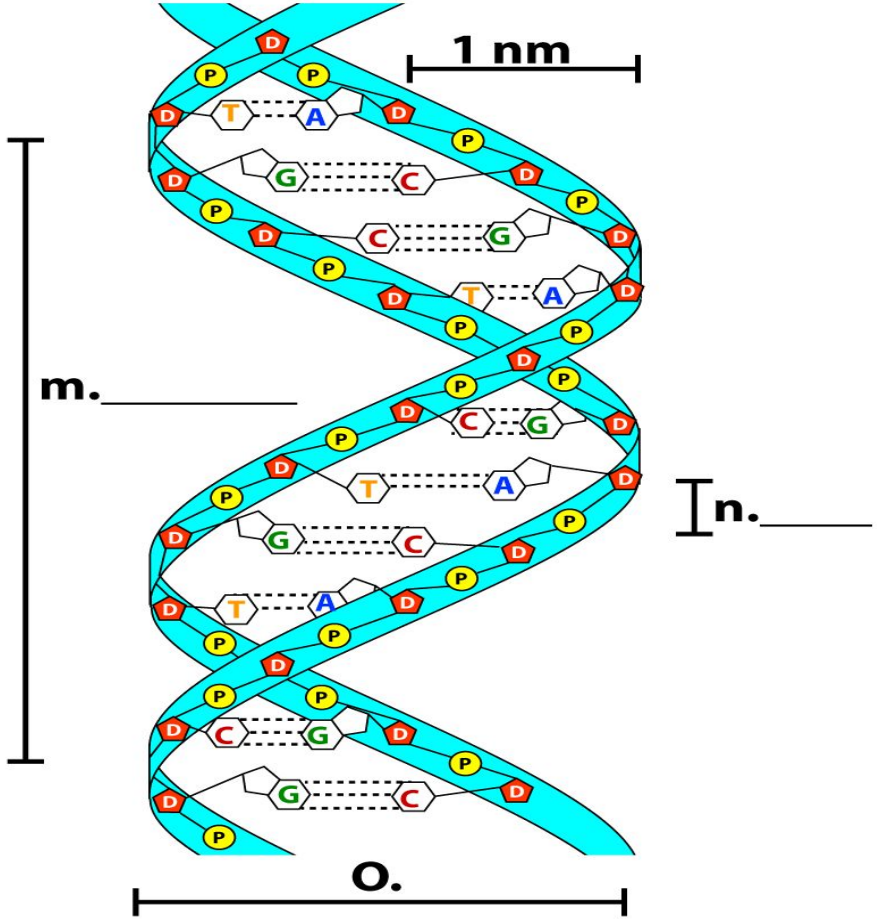
Difference b/w PAM and BLOSUM

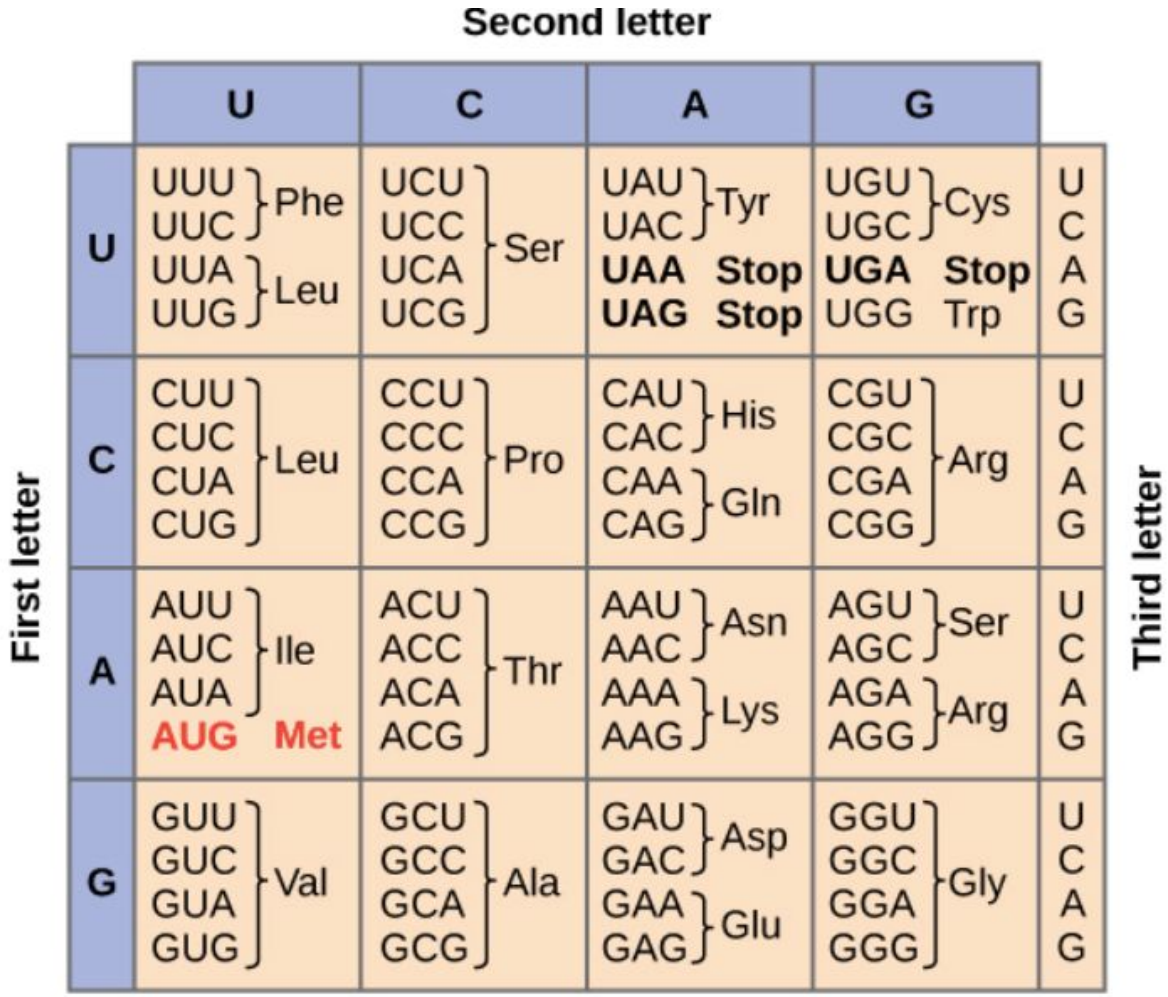
PAM	BLOSUM
Based on global alignments of closely related proteins.	Based on local alignments.
PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence but corresponds to 99% sequence identity.	BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.
Other PAM matrices are extrapolated from PAM1.	Based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
Higher numbers in matrices naming scheme denote larger evolutionary distance.	Larger numbers in matrices naming scheme denote higher sequence similarity and therefore smaller evolutionary distance.[19]


Sugar-phosphate backbone
P
D
G
T
C
A
Nitrogenous bases


P
D
T
A
C
G

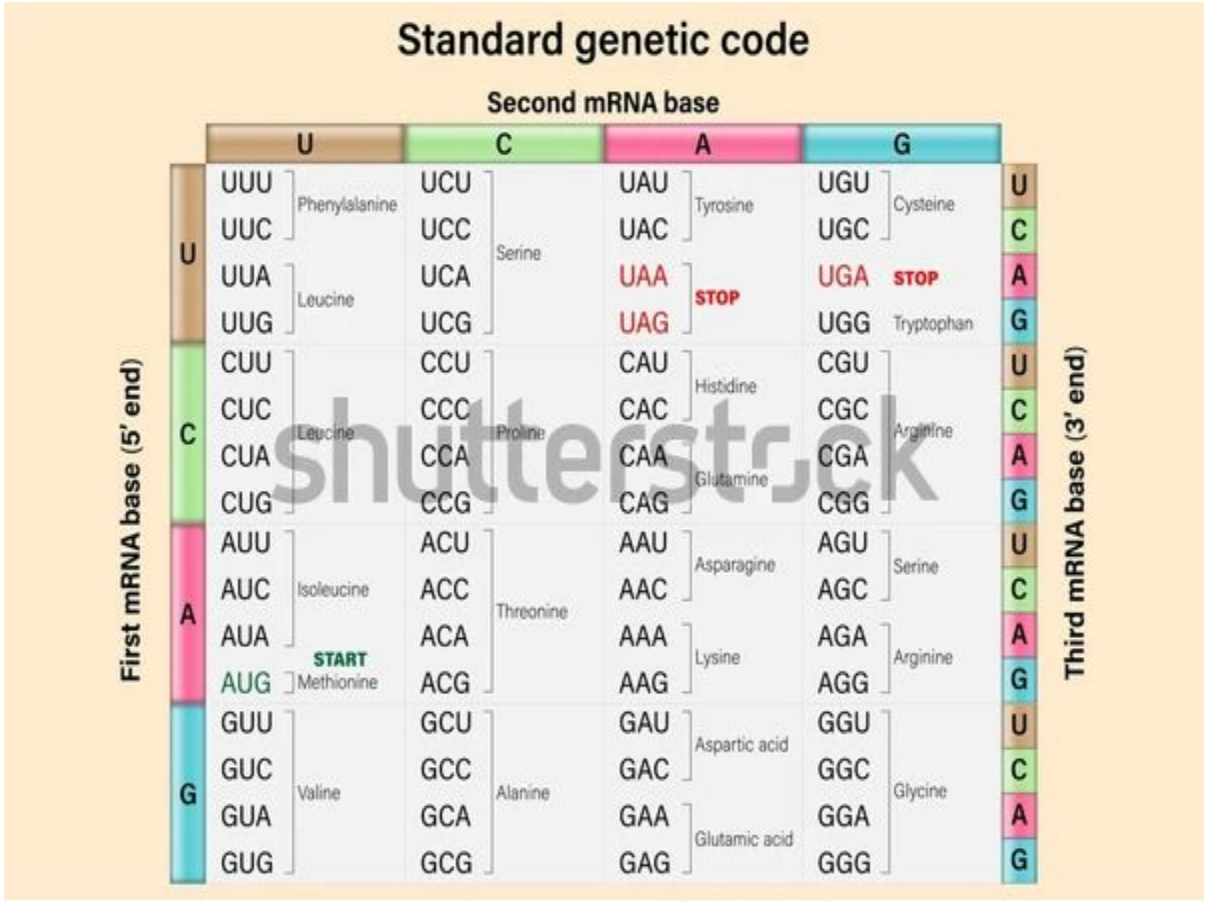





Second letter
U
C
A
G
First letter
Third letter
U
UUU Phe
UUC
UUA Leu
UUG
UCU Ser
UCC
UCA
UCG
UAU Tyr
UAC
UAA Stop
UAG Stop
UGU Cys
UGC
UGA Stop
UGG Trp
U C A G
C
CUU Leu
CUC
CUA
CUG
CCU Pro
CCC
CCA
CCG
CAU His
CAC
CAA Gln
CAG
CGU Arg
CGC
CGA
CGG
U C A G
A
AUU Ile
AUC
AUA
AUG Met
ACU Thr
ACC
ACA
ACG
AAU Asn
AAC
AAA Lys
AAG
AGU Ser
AGC
AGA Arg
AGG
U C A G
G
GUU Val
GUC
GUA
GUG
GCU Ala
GCC
GCA
GCG
GAU Asp
GAC
GAA Glu
GAG
GGU Gly
GGC
GGA
GGG
U C A G

mRNA Codon/Amino Acid Chart

First Base	Second Base				Third Base
	U	**C**	**A**	**G**	
U	UUU, UUC – Phenylalanine (Phe) UUA, UUG – Leucine (Leu)	UCU, UCC, UCA, UCG – Serine (Ser)	UAU, UAC – Tyrosine (Tyr) UAA, UAG – Stop	UGU, UGC – Cysteine (Cys) UGA – Stop UGG – Tryptophan (Trp)	U C A G
C	CUU, CUC, CUA, CUG – Leucine (Leu)	CCU, CCC, CCA, CCG – Proline (Pro)	CAU, CAC – Histidine (His) CAA, CAG – Glutamine (Glu)	CGU, CGC, CGA, CGG – Arginine (Arg)	U C A G
A	AUU, AUC, AUA – Isoleucine (Ile) AUG – Start Methionine (Met)	ACU, ACC, ACA, ACG – Threonine (Thr)	AAU, AAC – Asparagine (Asn) AAA, AAG – Lysine (Lys)	AGU, AGC – Serine (Ser) AGA, AGG – Arginine (Arg)	U C A G
G	GUU, GUC, GUA, GUG – Valine (Val)	GCU, GCC, GCA, GCG – Alanine (Ala)	GAU, GAC – Aspartic Acid (Asp) GAA, GAG – Glutamic Acid (Glu)	GGU, GGC, GGA, GGG – Glycine (Gly)	U C A G

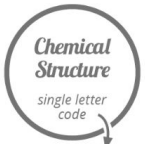


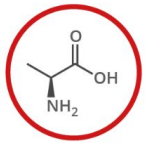
www.shutterstock.com · 2052705743

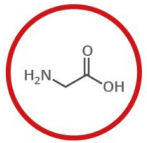
A GUIDE TO THE TWENTY COMMON AMINO ACIDS

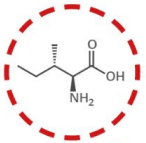
AMINO ACIDS ARE THE BUILDING BLOCKS OF PROTEINS IN LIVING ORGANISMS. THERE ARE OVER 500 AMINO ACIDS FOUND IN NATURE - HOWEVER, THE HUMAN GENETIC CODE ONLY DIRECTLY ENCODES 20. 'ESSENTIAL' AMINO ACIDS MUST BE OBTAINED FROM THE DIET, WHILST NON-ESSENTIAL AMINO ACIDS CAN BE SYNTHESISED IN THE BODY.

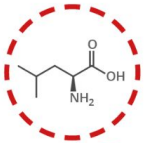
Chart Key: ALIPHATIC · AROMATIC · ACIDIC · BASIC · HYDROXYLIC · SULFUR-CONTAINING · AMIDIC · NON-ESSENTIAL · ESSENTIAL

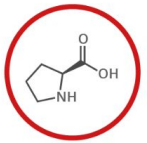


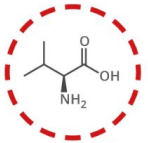

ALANINE A
ala

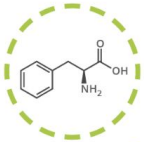

GLYCINE G
gly

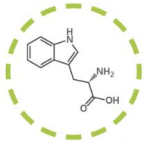

ISOLEUCINE I
ile

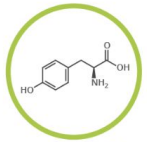

LEUCINE L
leu

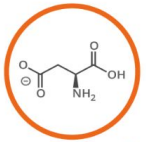

PROLINE P
pro

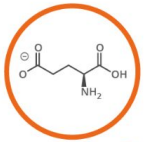

VALINE V
val

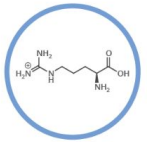

PHENYLALANINE F
phe

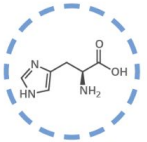

TRYPTOPHAN W
trp

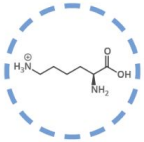

TYROSINE Y
tyr

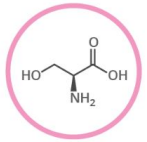

ASPARTIC ACID D
asp

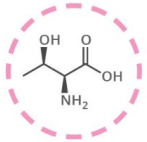

GLUTAMIC ACID E
glu

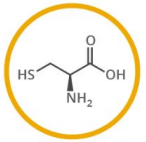

ARGININE R
arg

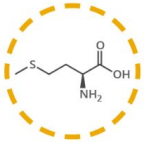

HISTIDINE H
his

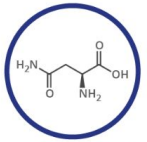

LYSINE K
lys

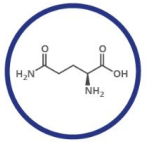

SERINE S
ser


THREONINE T
thr


CYSTEINE C
cys


METHIONINE M
met


ASPARAGINE N
asn


GLUTAMINE Q
gln

Note: This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between asparagine/aspartic acid and glutamine/glutamic acid is difficult. In these cases, the codes asx (B) and glx (Z) are respectively used.